Merkmalsbasiertes Opinion Mining anhand von Produktbewertungen in Onlineportalen

MASTERARBEIT

im Studiengang

Computer Science and Media

an der Hochschule der Medien in Stuttgart

vorgelegt von Christina Schneider

im Februar 2015

Erstprüfer: Prof. Dr.-Ing. Johannes Maucher, Hochschule der Medien, Stuttgart

Zweitprüfer: Prof. Walter Kriha, Hochschule der Medien, Stuttgart

Betreuer: Tobias Kässmann, Shopping24 GmbH, Hamburg

Erklärung

„Hiermit versichere ich, Christina Schneider, an Eides Statt, dass ich die vorliegende Masterarbeit mit dem Titel: „Merkmalsbasiertes Opinion Mining anhand von Produktbewertungen in Onlineportalen“ selbstständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden.
Ich habe die Bedeutung der eidesstattlichen Versicherung und die prüfungsrechtlichen Folgen (§ 19 Abs. 2 Master-SPO der HdM) sowie die strafrechtlichen Folgen (gem. § 156 StGB) einer unrichtigen oder unvollständigen eidesstattlichen Versicherung zur Kenntnis genommen.“

Stuttgart, am 2. Februar 2015

Christina Schneider

Kurzfassung

Die vorliegende Arbeit beschäftigt sich mit der Automatisierung der Merkmalsextraktion und der Stimmungsanalyse der extrahierten Merkmale in Produktbewertungen von Onlineportalen. Der Fokus liegt dabei auf der Merkmalsextraktion.
Ziel dieser Arbeit ist es zu eruieren, ob die automatisierte Extraktion der von Kunden bewerteten Merkmale eines Produktes aus einem Bewertungstext möglich ist. Ein weiteres Ziel ist es die von den Kunden geäußerte Meinung bezüglich der genannten Merkmale zu analysieren.
Mit dem Wissen aus dem Fazit dieser Arbeit soll es möglich sein eine Zusammenfassung der bewerteten Merkmale von Bewertungstexten sowie der zugehörigen Stimmungen zu erstellen. Eine solche Zusammenfassung dient dem Zweck sowohl potenziellen Käufern, als auch den Herstellern der Produkte einen Überblick über die geäußerten Stimmungen zu den erkannten Merkmalen eines Produktes zu verschaffen.
Es werden zwei Methoden zur Merkmalsextraktion angewendet und die Resultate anhand von Metriken verglichen. Die Merkmalsextraktion mit Hilfe von Word2Vec ist ein neuer Ansatz. Die Beschreibung und Evaluation dieses Ansatzes bilden den Hauptteil dieser Arbeit. Der zweite Ansatz ist die Extraktion der Merkmale mittels der häufigsten Nomen. Dessen Ergebnisse werden als Referenz für den neueren Ansatz herangezogen. Beide Methoden wurden bereits bei englischsprachigen Bewertungen erfolgreich eingesetzt. In dieser Arbeit werden beide Methoden auf deutschsprachige Bewertungen von Modeartikeln angewendet.
Zunächst werden Bewertungsdaten aus einem Onlineportal extrahiert und vorverarbeitet. Für die Merkmalsextraktion mittels häufigster Nomen werden die Daten Part-of-Speech annotiert und es wird diskutiert, wie aus den auf diese Weise gefundenen häufigsten Nomen tatsächliche Merkmale herausgefiltert werden können.
Bei Merkmalsextraktion mit Word2Vec wird eine manuell erstellte Liste von Merkmalen mit ähnlichen Begriffen erweitert.
Zum Ende der Arbeit wird eine Stimmungsanalyse beschrieben, welche anhand von Wortlisten mit positiver beziehungsweise negativer Konnotation die gefundenen Merkmale in positive, negative und neutrale Merkmale einteilt. Für die Merkmalsextraktion wird hierbei die Word2Vec Methode genutzt.
Sowohl die Ergebnisse der Merkmalsextraktion als auch der Stimmungsanalyse werden von zwei unabhängigen Nutzergruppen bewertet. Im Anschluss werden die Resultate der Nutzerbewertung diskutiert.

Abstract

This thesis addresses the automation of feature extraction and sentiment analysis on the extracted features of product reviews in online shops. The focus is on feature extraction.
The aim of this thesis is to investigate whether the automated extraction of product features, which have been rated by customers out of a review text, is possible. A further aim is to analyse the sentiment expressed by the customers regarding the aforementioned features.
With the knowledge gained from this work it should be possible to create a summary of the mentioned product features, as well as the associated sentiments. The purpose of such a summary is to give both the potential customers and the manufacturers of the products an overview of the expressed sentiments on the recognized features of a product.
Two methods are used for feature extraction and the results of both are compared using metrics. The feature extraction using Word2Vec is a new approach. The description and evaluation of this approach form the main part of this thesis. The second approach is the extraction of features by means of common nouns. The results of this approach are used as reference for the newer approach. Both methods have already been used successfully with english review texts. In this thesis both methods are applied to german text reviews of fashion items.
First the review data is extracted from an online shop and then preprocessed. For the feature extraction using the most common nouns, the data is part-of-speech annotated and it is discussed how actual features may be filtered from the so found common nouns.
For the feature extraction with Word2Vec a manually created list of features is extended with similar terms.
At the end of the thesis, a sentiment analysis approach is described, which divides the discovered features into positive, negative and neutral ones, based on word lists with positive or, respectively, negative connotations. Here, the Word2Vec feature extraction method is used.
Both the results of the feature extraction and the sentiment analysis are evaluated by two independent user groups. In the following the results of the user evaluations are discussed.

Inhaltsverzeichnis

Abkürzungsverzeichnis

CBOW Continuous bag-of-words

csv Comma-separated values

NLTK Natural Language Toolkit

POS Part-of-Speech

SKU Stock Keeping Unit

TnT Trigrams'n'Tags

URL Uniform Resource Locator

Kapitel 1

Einleitung

1.1 Motivation

Immer mehr Menschen verlassen sich bei Onlinekäufen nicht nur auf die Beschreibung des Produktes durch den Hersteller und den Verkäufer, sondern auch auf die Bewertungen durch andere Kunden.
Laut einer Studie von Bitcom [Bit12] liegt der Anteil an Onlinekäufern, welche vor dem Kauf Bewertungen anderer Käufer lesen, bei 73%. Gleichzeitig steigt aber auch die Anzahl der Bewertungen beliebter Produkte.
So wird es für Kunden immer schwieriger sich schnell einen Überblick über die von vorherigen Käufern geäußerten Meinungen zu verschaffen.
Bei einer großen Anzahl Bewertungen je Produkt ist es mühsam und oft zu zeitaufwändig alle Bewertungen zu lesen um sich ein umfassendes Bild von der Qualität und Beschaffenheit des Produktes zu machen.
Meist interessieren sich unterschiedliche Kunden auch für unterschiedliche Aspekte eines Produktes. Der eine Kunde achtet mehr auf den Tragekomfort einer Hose, der andere hat ein Auge auf den vorteilhaften Schnitt. Sucht man nun nach der Beschreibung eines bestimmten Merkmals in den Bewertungen, muss man meist viele davon lesen um sich umfassend über die Bewertungen zu diesem Merkmal zu informieren.
Abgesehen von den Kunden haben auch die Hersteller ein Interesse daran zu erfahren, was Kunden in den Bewertungen über ihr Produkt schreiben. Dieses Wissen kann zum Beispiel dafür genutzt werden das Produkt weiterzuentwickeln und zu verbessern. Es gibt auch einen Hinweis darauf, auf welche Merkmale die Kunden achten und wann sich monetäre Investitionen in die Weiterentwicklung der Produkte lohnen.
In dieser Arbeit werden daher die Bewertungen analysiert und programmatisch aufbereitet. Es wird mit Hilfe von Metriken bewertet, inwieweit es möglich ist aus Bewertungen die beschriebenen Merkmale sowie die jeweilige Stimmung zu diesen Merkmalen automatisiert und weitestgehend unüberwacht herauszufiltern.
Ist es möglich zuverlässig die Aspekte eines Produktes sowie die Stimmung dazu automatisiert aus den Bewertungen herauszulesen, so kann eine Zusammenfassung dieser Stimmungen erstellt werden.
Ziel ist für ein Produkt die bewerteten Merkmale und die jeweiligen Stimmungen zu filtern und eine Zusammenfassung der Merkmale und Stimmungen zu ermöglichen, so dass sich sowohl die potentiellen Käufer als auch die Hersteller einen Überblick über die bewerteten Merkmale und die Stimmungen verschaffen können.

1.2 Problemstellung

Da sich immer mehr Menschen bei Käufen im Internet auf die Meinungen anderer Kunden verlassen [Bit12], sollen diese Meinungen aufgeschlüsselt werden. Zumeist wird bei Bewertungen, die aus einer Sternebewertung von 1 (schlecht) bis 5 (sehr gut) Sternen und einer textuellen Bewertung in natürlicher Sprache bestehen, nur eine Zusammenfassung der Bewertungen aus dem Durchschnitt der Sternebewertung errechnet und dem Kunden angezeigt.
Mit dem Wissen aus dem Fazit dieser Arbeit soll es möglich sein zusätzlich zur vorhandenen Zusammenfassung der Sternebewertung eine Zusammenfassung der genannten Produktmerkmale in Bewertungstexten sowie der zugehörigen Stimmungen zu erstellen.
In dieser Arbeit soll untersucht werden inwieweit es möglich ist die einzelnen bewerteten Merkmale in den Bewertungen zu extrahieren. Es werden nur Merkmale betrachtet, welche aus einem Wort bestehen und explizit genannt werden.
Ebenso soll untersucht werden, ob es möglich ist die Stimmungen zu diesen Merkmalen zu erfassen. Dabei werden für die Merkmalsextraktion zunächst nur Bewertungen aus den Kategorien Damenmode und Herrenmode betrachtet.
Die Bewertungen selbst unterliegen in ihrer Anzahl einer ständigen Veränderung. Ebenso ändern sich die Produkte welche Bewertungen erhalten. Vor allem der Bereich der Mode ist sehr schnelllebig [BF10]. Ständig kommen neue Produkte hinzu und alte werden aus dem Sortiment genommen.
Die Extraktion der Merkmale und die Stimmungsanalyse müssen also ständig neu berechnet werden. Im besten Fall sollten nicht alle Bewertungen noch einmal durchlaufen werden müssen, sobald eine neue Bewertung für ein Produkt erstellt wurde.
Ein weiterer Aspekt der Bewertungen ist ihre Qualität. Die Texte sind von Kunden verfasst und daher nicht, wie bei redaktionell verfassten Texten, auf ihre Korrektheit geprüft.
Die Bewertungen sind oft umgangssprachlich verfasst, grammatikalisch nicht korrekt, enthalten Schreibfehler oder bestehen nicht aus vollständigen Sätzen. Häufig sind es nur Aufzählungen der positiven und negativen Aspekte des Produktes. Die Länge der Bewertungen variiert ebenfalls stark.
Daher wird betrachtet inwiefern eine Vorverarbeitung notwendig ist, um mit den Bewertungen gute Ergebnisse in der Merkmalsextraktion und der Stimmungsanalyse zu erzielen. Zur Evaluation der merkmalsbasierten Stimmungsanalyse wird neben den untersuchten Kategorien Damenmode und Herrenmode auch die Kategorie Multimedia herangezogen. Dabei soll gezeigt werden wie gut die Merkmalsextraktion bei einer fremden Domäne funktioniert.
Der Fokus der Arbeit liegt auf der Automatisierung der Merkmalsextraktion, insbesondere der Merkmalsextraktion mit dem Word2Vec Algorithmus. Die Merkmalsextraktion mit Hilfe der häufigsten Nomen wird dabei als Referenz herangezogen um einen Vergleich der Ergebnisse zu ermöglichen.

1.3 Aufbau der Arbeit

Die Vorgehensweise und damit der Aufbau der Arbeit ergibt sich aus dem allgemeinen Prozess des Data Mining (siehe Abbildung 1.1). Die Auswahl der Daten sowie die Vorverarbeitung werden in Kapitel 2 beschrieben. Die Kapitel 3 und 4 umfassen Transformation sowie das Finden der Muster. Die anschließende Evaluation findet in Kapitel 5 statt.
Mit dem daraus entstandenen Wissen kann dann in Kapitel 6 und 7 die merkmalsbasierte Stimmungsanalyse realisiert und evaluiert werden. Die in Abbildung 1.2 dar-

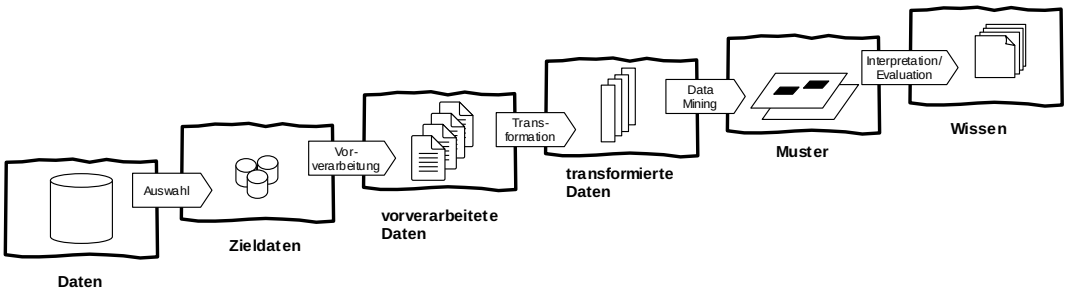

Abbildung 1.1: Prozess des Data Mining (nach [FPS96])

gestellte Vorgehensweise spiegelt sich auch in der Anordnung der Kapitel wider, welche im Folgenden kurz vorgestellt werden. Neben dem Einleitungskapitel ist die vorliegende Arbeit in 7 Teile gegliedert. Neben den Kapiteln wird zusätzlich kurz auf die Dokumente im Anhang und den Inhalt der beiliegenden CD eingegangen.

Kapitel 2 - Datengrundlage und Vorverarbeitung
Hier wird erläutert, welche Daten für diese Arbeit verwendet wurden. Außerdem wird dargestellt welche Schritte der Vorverarbeitung diese Daten durchlaufen müssen um sie für die nachfolgenden Arbeitsschritte verwenden zu können.

Kapitel 3 - Merkmalsextraktion durch häufigste Nomen
In Kapitel 3 wird die erste Methode zur Extraktion von Merkmalen vorgestellt und die daraus entstandenen Ergebnisse evaluiert.

Kapitel 4 - Merkmalsextraktion mit Word2Vec
Kapitel 4 stellt den Hauptteil der Arbeit dar. Hier wird die Merkmalsextraktion mit Hilfe von Word2Vec Modellen vorgestellt. Die einzelnen Arbeitsschritte werden erläutert und bewertet. Die Ergebnisse, die in diesem Kapitel gewonnen wurden dienen als Basis für das Opinion Mining, welches in Kapitel 6 behandelt wird.

Kapitel 5 - Vergleich der Methoden zur Merkmalsextraktion
In diesem Kapitel werden die vorgestellten Methoden zur Merkmalsextraktion verglichen und ihre Eignung im Hinblick auf den vorliegenden Kontext bewertet.

Kapitel 6 - Merkmalsbasiertes Opinion Mining
Wie oben erwähnt behandelt das 6. Kapitel das Opinion Mining. Das dazu verwendete System wird erklärt und die Vorgehensweise erläutert.

Kapitel 7 - Bewertung der Ergebnisse aus dem merkmalsbasierten Opinion Mining
Die Herangehensweise der Evaluation der Ergebnisse des in Kapitel 6 erläuterten Systems wird in diesem Kapitel erklärt und die Ergebnisse betrachtet.

Kapitel 8 - Fazit und Ausblick
Im letzten Kapitel werden die Ergebnisse dieser Arbeit zusammengefasst und es wird festgestellt, inwieweit die Anforderungen aus der Problemstellung erfüllt wurden.
Zusätzlich gibt das Kapitel einen Ausblick auf mögliche weiterführende Arbeiten, die auf den Erkenntnissen dieser Arbeit aufbauen können.

Anhang A
Der Anhang A ist dem Kapitel 2 zugeordnet und beinhaltet einen Überblick über das Stuttgart Tübingen Tagset, das Perl Script zur Bereinigung der Wikipediadaten sowie die Änderungen welche daran vorgenommen wurden.

Anhang B
Die Dokumente im Anhang B gehören zu den Kapiteln 3 und 4. Es handelt sich um die häufigsten Nomen aus der Merkmalsextraktion in Kapitel 3 sowie die Liste der Wortpaare, die zum Test der Modelle in Kapitel 4 verwendet wurden. Ebenfalls in Anhang B befindet sich die Liste der manuell annotierten Merkmale sowie die unterschiedlichen Versionen der Schreibweise des Wortes *Preisleistungsverhältnis*.

CD
Auf der CD befindet sich die vorliegende Arbeit im PDF Format und der Sourcecode aller für diese Arbeit erstellten Programme.

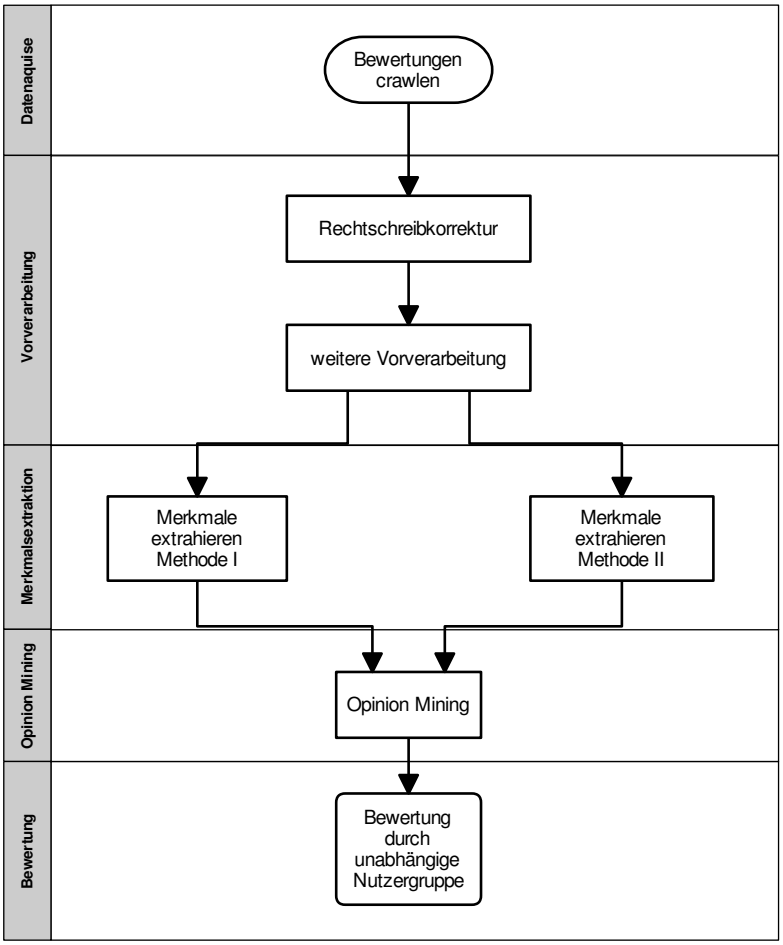

Abbildung 1.2: Ablauf

1.4 Verwandte Arbeiten

In den letzten Jahren gab es immer mehr Bemühungen nutzergenerierte Inhalte automatisiert zu verarbeiten und Wissen daraus zu generieren. Dabei waren die vorrangigen Ziele Produktmerkmale aus Texten zu extrahieren und die Stimmungen zu diesen Merkmalen zusammenzufassen.
Nachdem Arbeiten zur Stimmungsanalyse schon gute Ergebnisse erzielen, wird immer mehr auch der Fokus auf die Merkmalsextraktion und die merkmalsbasierte Stimmungsanalyse gesetzt. Arbeiten in diesem Bereich, welche sich nicht nur mit der Stimmungsanalyse, sondern auch mit der Merkmalsextraktion befassen, lassen sich grob in folgende Kategorien einteilen:

(1) Es wird mit einem manuell erstellten Set von Merkmalen gearbeitet. Der Fokus liegt auf der Sentimentanalyse, also der Filterung der Stimmungen je definiertem Merkmal [DLY08; Pon12]. Diese Methode ist sehr zeitaufwändig und erfordert spezielles Wissen über die Domäne um die Liste der Merkmale zu erstellen. Um diese Methode in einer realen Anwendung nutzen zu können muss die Liste ständig redaktionell überprüft und verbessert beziehungsweise erweitert werden.
Hinzu kommt das Problem, dass Kunden auch nicht vorhandene Merkmale eines Produktes bewerten beziehungsweise das Fehlen derselben bemängeln. Diese Merkmale sind nicht immer offensichtlich und daher bei einem manuell erstellten Set nur schwer zu erfassen. Gerade jedoch fehlende Merkmale sind für die Hersteller eines Produktes für die Weiterentwicklung besonders interessant.

(2) Statt mit einem vordefinierten Set von Merkmalen zu arbeiten wird mit Hilfe eines vorhergehenden Part-of-Speech Tagging eine Statistik der Häufigkeit aller Nomen erstellt. Danach werden manuell die Merkmale aus den häufigsten Nomen gewählt und es wird mit diesen gearbeitet [Kim10; Ton11]. Hier ist die Merkmalsextraktion schon zum Teil automatisiert und bedarf weniger redaktioneller Nacharbeitung. Es ist jedoch immer noch ein manueller Eingriff notwendig, bei dem ebenfalls domänenspezifisches Wissen erforderlich ist.
Diese Methode funktioniert bei einer statischen Menge an Bewertungen schon recht gut. Jedoch muss auch hier, sobald neue Produkte hinzukommen oder die Anzahl der Bewertungen sich ändert, manuell eingegriffen werden und der Schwellwert, ab welcher Häufigkeit ein Wort potenziell ein Merkmal darstellt, entsprechend angepasst werden. Dieser Ansatz beruht auf der Annahme, dass nur Nomen oder Nominalphrasen Merkmale darstellen. Das Part-of-Speech Tagging läuft automatisiert ab, jedoch müssen die tatsächlichen Merkmale am Ende manuell aus der Liste der häufigsten Nomen gewählt werden.
Mit dieser Vorgehensweise können nur häufig auftretende Merkmale gefunden werden.

(3) Neuere Arbeiten fokussieren stärker auf die zunehmende Automatisierung der Merkmalsextraktion.
Teils wird hier wie beim zweiten Ansatz auf die Häufigkeit der Nomen zurückgegriffen. Hierbei werden jedoch nicht die Merkmale anhand der Häufigkeiten von Hand selektiert, sondern ein Schwellwert bestimmt, ab welcher Häufigkeit ein Wort als Merkmal definiert wird [Jor10]. Bei dieser Methode werden bei niedrigem Schwellwert sehr viele

Nomen als Merkmale deklariert, welche keine Merkmale sind. Bei hohem Schwellwert werden viele tatsächliche Merkmale nicht als Merkmale erkannt, da sie nicht häufig genug in den Trainingsdaten vorkommen.
Mit der in [HL04] vorgestellten Methode werden nicht nur Nomen sondern auch Nominalphrasen als Merkmale erkannt. Auch hierzu werden die Annotationen aus dem Part-of-Speech Tagging herangezogen. Worte aus den Nominalphrasen, welche häufig zusammen in einem Satz vorkommen und innerhalb eines Satzes einen möglichst geringen Abstand zueinander aufweisen, werden ebenfalls als Merkmale definiert.
Alternativ werden zusätzlich zur Worthäufigkeit Aspekte wie syntaktische Muster betrachtet. Als vielversprechend hat sich das Muster „Nomen, welches einem Adjektiv folgt“ erwiesen [Bla+08]. Hier werden nach Anwendung dieses Musters die so gefundenen Merkmale noch einmal gefiltert. Merkmale, welche aus Stopworten bestehen oder selten vorkommen, werden herausgefiltert. Außerdem werden Merkmale, die nicht oder selten in Zusammenhang mit einem Meinungswort genannt werden, ebenfalls nicht weiter betrachtet.
Bei [Bla+08] wird auch der Ansatz genannt weiteres Wissen außerhalb des Textes zur Merkmalsextraktion zu nutzen, wie die Information aus welcher Domäne der Text stammt. Je nach Domäne, also dem Kontext aus welchem der Text stammt, werden unterschiedliche Begriffe und Worte genutzt. Auch, ob in einer Domäne eher umgangssprachliche oder redaktionell erstellte Texte vorkommen, ist hierbei von Interesse.
In [KV12] wird ebenfalls ein regelbasierter Ansatz für die Merkmalsextraktion genutzt. Hier wird mit einem domänenspezifischen Verzeichnis von Ausdrücken gearbeitet.
In [PE07] werden nicht nur explizite, sondern auch implizite Merkmale identifiziert (zur Definition expliziter und impliziter Merkmale siehe Kapitel 2.1.1). Dies wird erreicht, indem Meinungsworte, die explizite Merkmale beschreiben, in Cluster eingeteilt werden. Wird eine Meinung über ein implizites Merkmal geäußert, so wird anhand des genutzten Meinungswortes diese über das zugehörige Cluster einem expliziten Merkmal zugeordnet.
Sind nur wenige Bewertungen vorhanden oder kommen die Bewertungen aus sich in ihren Merkmalen stark unterscheidenden Domänen sind alle Vorgehensweisen der vorgestellten drei Kategorien nur bedingt geeignet.
Ein Ansatz, der dazu dienen soll seltene Merkmale zu finden, wird in [HL04] vorgestellt. Hierbei werden zuerst die Meinungsworte gesucht und danach die Merkmale bestimmt, zu denen das Meinungswort zugeordnet werden kann. Dieser Ansatz wird jedoch nur in Ausnahmefällen verwendet und zwar, wenn durch die Häufigkeitsanalyse aller möglichen Merkmale nicht alle tatsächlichen Merkmale gefunden werden konnten.

Ein neuer Ansatz, der nicht auf der Häufigkeit von Nomen beruht, wurde beim internationalen Workshop für semantische Evaluation 2014 (SemEval-2014)[1] eingereicht. Dort wurde eine Aufgabe zur automatischen Merkmalsextraktion und Zuordnung der Meinungen zu den Merkmalen gestellt [Pon+14]. Das Team, welches bei der Merkmalsextraktion aus den Berwertungstexten der zwei dort betrachteten Domänen die besten beziehungsweise zweitbesten Ergebnisse erzielte, nutzt unter anderem K-Means Cluster errechnet aus Word2Vec Modellen, welche mit Bewertungsdaten von Amazon und Yelp trainiert wurden [TW14].
Hierbei wurden auch Mehrwort Merkmale betrachtet. Der Vorteil dieses Ansatzes ist,

[1] http://alt.qcri.org/semeval2014/, abgerufen am 19.11.2014

dass hier kein Part-of-Speech Tagging notwendig ist. Alle vorherigen Ansätze stützen sich darauf, dass die Genauigkeit bei der Annotation der Worte mit Part-of-Speech Tags hoch ist. Können die Texte aufgrund ihrer Qualität nicht genau genug Part-of-Speech annotiert werden, sinkt auch die Qualität der Merkmalsextraktion.

In dieser Arbeit wird ein neuer Ansatz vorgestellt, welcher wie bei [TW14] Word2Vec Modelle nutzt, um Merkmale zu extrahieren. Zusätzlich dazu wird eine Merkmalsextraktion mittels der häufigsten Nomen durchgeführt. Die Ergebnisse der Extraktion der Merkmale mittels der häufigsten Nomen dienen als Referenz für den neuen Ansatz.

Kapitel 2

Datengrundlage und Vorverarbeitung

Im Kapitel *Datengrundlage und Vorverarbeitung* wird erläutert, wie die für die Merkmalsextraktion genutzten Daten gewonnen wurden. Außerdem werden die Vorverarbeitungsschritte beschrieben, welche durchlaufen werden müssen, bevor die Rohdaten für die Analysen genutzt werden können. Die Reihenfolge der einzelnen Schritte ist auch in Abbildung 2.1 dargestellt.
Nach dem Prozess des Data Mining (Abbildung 1.1) werden in diesem Abschnitt die Zieldaten ausgewählt und vorverarbeitet.

2.1 Definitionen

Im folgenden Abschnitt wird definiert, was die verwendeten Begriffe im Kontext dieser Thesis bedeuten.

2.1.1 Merkmale

Als Merkmale werden hier Bestandteile oder Eigenschaften der bewerteten Produkte bezeichnet. Das Produkt selbst ist kein Merkmal. Daher werden allgemeine Bewertungen eines Produktes nicht in der Sentimentanalyse betrachtet.
Beispiele für Merkmale der Produkte aus dem Bereich Mode sind *Kragen* als Bestandteil eines Kleidungsstückes und *Tragekomfort* als Eigenschaft eines Modeartikels.
Es werden nur explizite Merkmale, welche als Nomen vorkommen, betrachtet.

Es gibt explizite und implizite Merkmale [DLY08]. Der Unterschied wird im Folgenden kurz dargestellt.
Ein Beispiel für ein explizites Merkmal ist:

Die Qualität des T-Shirts war schlecht.

Hier ist *Qualität* das Merkmal, welches bewertet wird.
Implizite Merkmale sind, selbst für einen menschlichen Tagger, schwerer zu identifizieren und werden daher hier nicht betrachtet. Ein Beispiel für ein implizites Merkmal ist:

Das T-Shirt hat sich nach der ersten Wäsche verzogen.

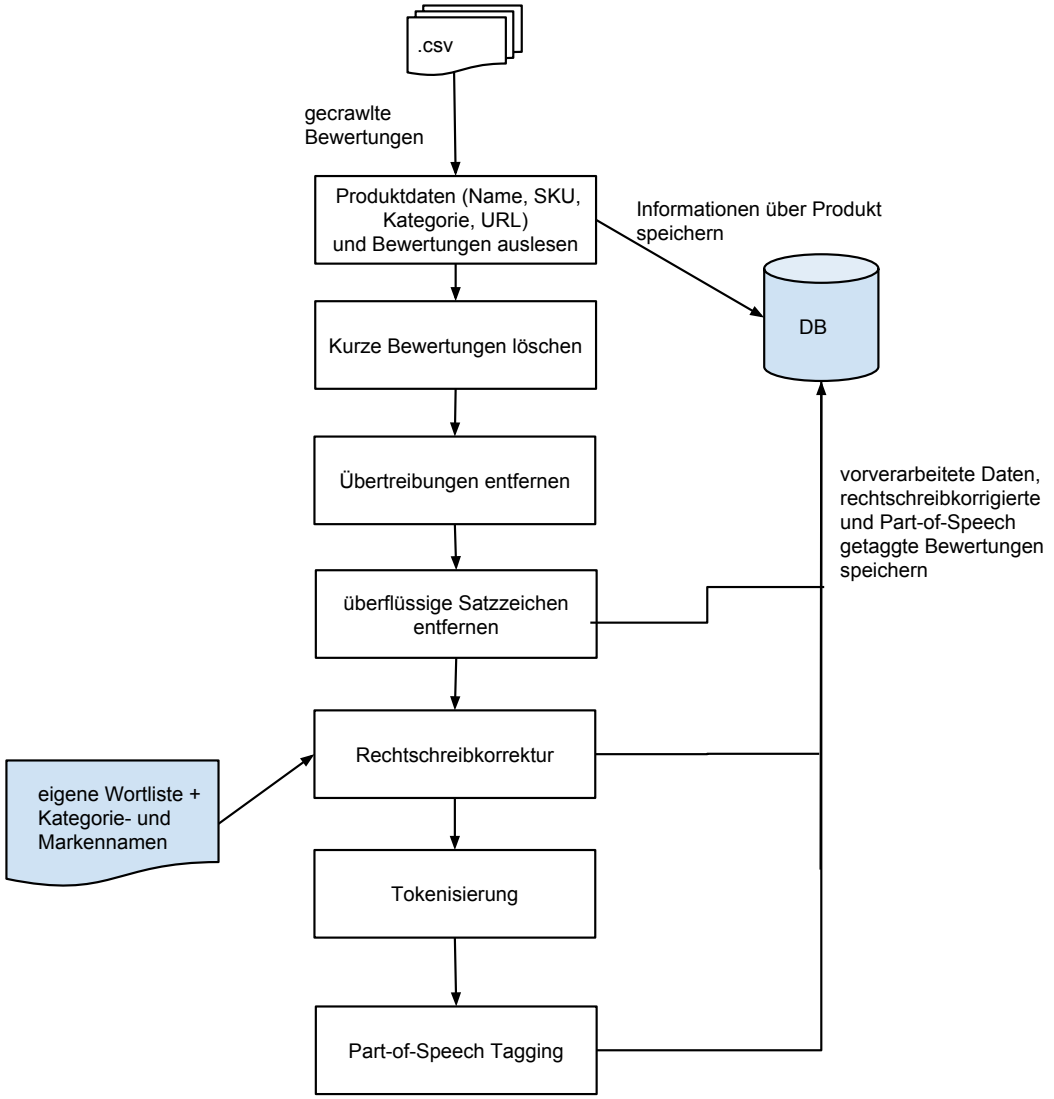

Abbildung 2.1: Ablaufdiagramm zum Vorverarbeitungsprozess

Hierbei wird ebenfalls die Qualität des Artikels bewertet. Jedoch ist dies nicht auf den ersten Blick ersichtlich und daher schwer automatisch zu filtern.

2.1.2 Meinungen

Als Meinungen wird hier der Kontext bezeichnet, in welchem ein Merkmal auftaucht. Dies kann ein positiver oder negativer Kontext sein.
Diese Meinungen werden pro Bewertung für jedes Merkmal zusammengefasst und anhand eines numerischen Wertes dargestellt. In dieser Arbeit wird nur eine positive oder negative Tendenz angegeben, wobei +1 für eine positive und -1 für negative Tendenz der Meinung steht.

Meinungsworte

Als Meinungsworte werden hier Worte bezeichnet, welche eine positive oder negative Meinung zu einem bestimmten Merkmal ausdrücken.
Die Meinungsworte und ihre Stimmung sind aus den von Hand annotierten Bewertungen (siehe Kapitel 4.4) entnommen. Zusätzlich wurde eine Liste von Meinungsworten aus dem SentiWS genutzt (siehe Kapitel 6), einer Sammlung von deutschen Worten mit positiver oder negativer Konnotation.
Meinungsworte können Nomen, Adjektive, Verben oder Adverbien sein [RQH10].

2.2 Datenbasis

Es wurden Produktbewertungen von Artikeln der Bereiche Damenmode, Herrenmode und Multimedia des Onlineportals *otto.de*[1] verwendet. Diese sind von den Kunden des Portals verfasst und nicht auf Rechtschreibfehler überprüft. Der Hauptteil der Bewertungen ist in deutscher Sprache verfasst.
Um eine möglichst große Datenbasis zu erhalten wurden nur Produkte berücksichtigt, welche mehr als 100 Bewertungen aufweisen. Dabei wurden für jedes dieser Produkte die jeweiligen Bewertungen in Textform sowie je Bewertung die Sternebewertungen im Bereich von 1-5 Sternen, die Stock Keeping Unit (SKU) des bewerteten Produktes, dessen Namen, die Kategorie, sowie die URL gespeichert.

2.3 Externe Programme und Hilfsmittel

In diesem Abschnitt erfolgt eine Auflistung, sowie eine kurze Beschreibung der externen Programme und anderer Hilfsmittel, welche für diese Arbeit verwendet wurden.

Scrapy ist ein Open Source Python Framework[2], welches für schnelles scraping und crawling von Webseiten verwendet wird. Es lassen sich hiermit einfach und ohne großen Aufwand Crawler erstellen.
Ein Crawler ist ein Programm, mit welchem Webseiten durchsucht werden können. Als scraping wird dabei der Vorgang bezeichnet den Inhalt der Webseite auszulesen.

STTS[3] ist das Stuttgart-Tübingen-Tagset, welches für die deutsche Sprache das bekannteste Tagset darstellt. Eine Übersicht der Tags findet sich im Anhang A.1.

pyenchant ist eine Bibliothek[4] zur Rechtschreibkorrektur für Python, basierend auf der *Enchant* Bibliothek[5].

Hunspell[6] ist eine Software zur Rechtschreibkorrektur. Sie kann eigenständig genutzt werden. Die einzelnen Wörterbücher für verschiedene Sprachen können aber auch in

[1] http://www.otto.de
[2] scrapy.org
[3] http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html, abgerufen am 25.01.2015
[4] http://pythonhosted.org/pyenchant/, abgerufen am 02.10.2014
[5] http://abisource.com/projects/enchant/, abgerufen am 02.10.2014
[6] http://hunspell.sourceforge.net/, abgerufen am 25.01.2015

Verbindung mit *pyenchant* verwendet werden.

NLTK oder auch Natural Language Toolkit[7] [BKL09] ist eine Sammlung von Hilfsmitteln zur Verarbeitung natürlicher Sprache in Python Programmen. Es beinhaltet unter anderem Bibliotheken für Segmentierung in Sätze und Worte und das Part-of-Speech Tagging (siehe Kapitel 3.1).

PostgreSQL[8] ist ein Open Source Datenbankmanagementsystem.

psycopg [9] ist ein *PostgreSQL* Adapter für Python, der es erlaubt aus einem Python-Programm heraus Datenbankoperationen auf einer *PostgreSQL* Datenbank auszuführen.

gensim[ŘS10][10] ist eine Python Bibliothek. Sie enthält unter anderem eine performanzoptimierte Implementierung von Word2Vec.

brat[11] rapid annotiation tool ist ein browserbasiertes Programm um Texte schnell und einfach zu annotieren. Die Annotationen können dabei je nach Anwendungsfall individuell definiert werden.

2.4 Crawling

Die Produktbewertungen wurden mit Hilfe eines Crawlers vom Online-Shoppingportal otto.de[12] gesammelt. Die für das Crawling verwendeten Kategorien sind die Hauptkategorien Damenmode, Herrenmode und Multimedia sowie alle ihre untergeordneten Kategorien.
In den folgenden Kapiteln wird mit den Kategorien Damenmode und Herrenmode gearbeitet und am Ende die merkmalsbasierte Stimmungsanalyse zusätzlich mit der Kategorie Multimedia getestet.

2.4.1 Vorgehensweise

Das Auslesen von Daten aus Webseiten erfolgt mit Hilfe eines Bots. Dieser durchsucht eine gegebene URL und folgt den Hyperlinks bis zu einer definierten Tiefe (Crawling). Ist eine Produktseite erreicht werden bestimmte Daten der Seite ausgelesen (Scraping) und gespeichert. Eine Produktseite wird anhand ihrer URL erkannt, welche folgendermaßen aufgebaut ist: *http://www.otto.de/p/*[Produktname]-[SKU].
Für das Auslesen der Daten aus einer Seite wird anhand der HTML-Struktur, welche für alle Produktseiten gleich ist, festgelegt, welcher Teil der Seite die Bewertungen enthält. Dieser Teil wird persistent gespeichert.

Der Crawler wurde mit Hilfe des Python Frameworks *Scrapy* geschrieben (siehe Kapitel

[7] www.nltk.org, abgerufen am 02.10.2014
[8] www.postgresql.org, abgerufen am 15.10.2014
[9] initd.org/pyscopg/, abgerufen am 14.10.2014
[10] https://radimrehurek.com/gensim/, abgerufen am 11.01.2015
[11] http://brat.nlplab.org, abgerufen am 02.01.2015
[12] http://www.otto.de

2.3). Für das Crawling wurde eine Liste der URLs der Unterkategorien der Hauptkategorien Damenmode, Herrenmode und Multimedia erzeugt. Diese wurden vom Crawler bis auf Produktebene verfolgt.
Für jedes hierbei gefundene Produkt, welches mehr als 100 Bewertungen aufweist, wurden die Bewertungen ausgelesen und gespeichert. Für jede Unterkategorie wurde eine .csv Datei erstellt, welche alle Bewertungen enthält. Im Detail wurden hierbei für jedes Produkt folgende Informationen gespeichert:

- SKU
- URL
- Kategorie (inklusive Elternkategorien)
- Name des Produktes
- Bewertungen

Die Bewertungen an sich enthalten wiederum unter anderem den Bewertungstext, sowie die allgemeine Bewertung, welche durch maximal 5 vergebene Sterne ausgedrückt wird.

2.5 Analyse der Trainingsdaten

Vor der Vorverarbeitung der Daten wurden diese analysiert um einen Überblick über die Menge und Beschaffenheit der Bewertungen zu erhalten.

Tabelle 2.1: Anteil der Produkte mit mehr als 100 Bewertungen an der Gesamtmenge aller Produkte der Stichprobe (gerundete Werte)

Kategorie	Anzahl	Anteil
Mode Damen + Herren	131 000	$3,7\%$
Mode Damen	86 000	$4,5\%$
Mode Herren	45 000	$2,1\%$

Tabelle 2.1 zeigt die Menge aller Produkte, welche zum Zeitpunkt des Crawlings in den Kategorien Damen und Herren zur Verfügung standen. Zusätzlich zeigt sie den Anteil der Produkte, welche mehr als 100 Bewertungen haben. Im Schnitt haben demnach $3,7\%$ der Produkte aus den Kategorien Damen und Herren mehr als 100 Bewertungen. Dabei haben im Bereich Damen im Schnitt deutlich mehr Produkte viele Bewertungen als in der Kategorie Herren. Anzumerken ist, dass sowohl in der Gesamtanzahl als auch im Anteil Produkte enthalten sind, welche in mehreren Unterkategorien vorkommen.

Tabelle 2.2: Anzahl der Produkte mit mehr als 100 Bewertungen, Menge an Bewertungen sowie durchschnittliche Wortanzahl je Bewertung

Kategorie	Anzahl	Bewertungen	ØWortanzahl
Mode Damen + Herren	3008	793519	$17,9$
Mode Damen	2436	639085	$18,5$
Mode Herren	572	154434	$15,2$

Im Gegensatz zu Tabelle 2.1 wurde in Tabelle 2.2 jedes Produkt nur einmal berücksichtigt. Sie zeigt die Gesamtanzahl der gecrawlten Bewertungen sowie deren durchschnitt-

liche Länge, dargestellt durch die Wortanzahl.
Wie in Tabelle 2.1 bereits gezeigt enthält die Kategorie Damen im Schnitt deutlich mehr häufig bewertete Produkte als der Bereich Herren. Zusätzlich sind die Bewertungen in der Kategorie Damen durchschnittlich auch länger. Insgesamt wurden über 3000 Produkte mit insgesamt fast 800 000 Bewertungen gecrawlt.

2.6 Vorverarbeitung

Da es sich bei Bewertungen in Onlineportalen um Texte handelt, welche nicht redaktionell verfasst sind, sondern von den Nutzern des Onlineportals geschrieben wurden, sind diese oft ohne Vorverarbeitung nur eingeschränkt verwertbar.
Daher werden die Bewertungen zunächst vorverarbeitet. Dazu gehört die Entfernung von sehr kurzen Bewertungen und Übertreibungen, die Rechtschreibkorrektur, die Segmentierung in Sätze und Worte, sowie das Part-of-Speech Tagging.

2.6.1 Übertreibungen und kurze Bewertungen

Um die Daten für die Rechtschreibkorrektur vorzubereiten, werden zunächst Übertreibungen korrigiert. Zusätzlich dazu werden Bewertungen, welche aus weniger als 10 Zeichen bestehen gelöscht, da diese meist nur aus einem Wort wie zum Beispiel „Super!“ bestehen und somit für die Merkmalsextraktion irrelevant sind.
Zu Übertreibungen gehören mehrfach auftretende Satzzeichen hintereinander wie zum Beispiel „Artikel xy ist super!!!!“. Zu Übertreibungen gehören ebenfalls langgezogene Worte wie in „Artikel xy is suuuuuper!“.
Zur Korrektur der Übertreibungen werden von mehreren gleichen, direkt aufeinanderfolgenden Satzzeichen alle bis auf eines entfernt. Bei mehrfach gleichen, direkt hintereinander auftretenden Buchstaben werden alle bis auf zwei entfernt. Dadurch kann mit Hilfe der folgenden Rechtschreibkorrektur je nach Fall ein fehlender Buchstabe ergänzt, beziehungsweise ein weiterer überflüssiger entfernt werden.

2.6.2 Rechtschreibkorrektur

Wird vor der Merkmalsextraktion eine Korrektur von Rechtschreibfehlern durchgeführt, kann damit eine Verbesserung der Ergebnisse der nachfolgenden Schritte erreicht werden. Selbst simple Fehler, welche durch eine Rechtschreibprüfung leicht korrigiert werden können, führen sonst zu einer Einschränkung der Verwertbarkeit der verwendeten Texte [Car+09, Kap. 5.1]. Zur Rechtschreibkorrektur wird *pyenchant* verwendet (siehe Kapitel 2.3). Es wurde das deutsche Wörterbuch der Rechtschreibprüfung Hunspell[13] verwendet, welches unter anderem auch in LibreOffice verwendet wird. Zusätzlich zum Hunspell Wörterbuch wurden zur Korrektur eine Liste von Markennamen und Kategorienamen genutzt. Dies ist mit *pyenchant* sehr einfach möglich. Es wird eine Textdatei angelegt, in der in jeder Zeile ein Wort steht. Der Pfad zu dieser Textdatei wird in *pyenchant* zusätzlich zum Wörterbuch angegeben.
Jedes Wort einer Bewertung, welches nicht im Wörterbuch vorkommt und länger als ein

[13] http://hunspell.sourceforge.net/, abgerufen am 19.01.2015

Zeichen ist, wird durch den ersten Eintrag der Liste an Korrekturvorschlägen ersetzt, sofern eine solche verfügbar ist.
Dieser Ansatz führt dazu, dass Worte, die zwar richtig sind, aber nicht im Wörterbuch vorkommen, durch ein ähnliches Wort ersetzt werden. Um die dadurch entstehenden Fehler gering zu halten wurde die Wortliste um Worte ergänzt, die häufig in den Bewertungen auftauchen, jedoch nicht im Wörterbuch enthalten sind. Beispiele dafür sind germanisierte englische Begriffe wie *stylisch* und *Teenie*, aber auch umgangssprachliche Ausdrücke wie *fusselt* und *schick*.
Da *enchant* für eine interaktive Rechtschreibkorrektur in Programmen zur Textverarbeitung gedacht ist, wird für ein Wort, welches nicht im Wörterbuch vorkommt, zwar eine Liste an Korrekturvorschlägen gegeben, jedoch kein Parameter mit welcher Sicherheit diese auch zutreffen.
Ein Weg die Fehler durch die Rechtschreibkorrektur zu mindern wäre es ein reduziertes Wörterbuch zu erstellen, welches nur noch Worte aus der Domäne enthält. Dies ist jedoch sehr zeitaufwändig und erfordert spezielles Wissen über die Domäne.

2.6.3 Satzsegmentierung

Um die Bewertungen in Sätze zu unterteilen wurde NLTK (siehe Kapitel 2.3) verwendet. Hier gibt es Modelle zur Satzsegmentierung, welche speziell auf bestimmte Sprachen zugeschnitten und trainiert wurden. Diese Spezialisierung ist notwendig um zu verhindern, dass nach Abkürzungen ein Satz fälschlicherweise getrennt wird [BKL09, Kap. 3.8].
Bei nutzergenerierten Inhalten kommen zusätzliche Probleme bei der Satzsegmentierung hinzu. Die Bewertungen sind oft nur in Kleinbuchstaben geschrieben und teils ist die Interpunktion falsch gesetzt oder fehlt völlig. Einige Bewertungen bestehen auch nur aus stichwortartigen Aufzählungen.
Trotz der zusätzlichen Herausforderungen, welche die nutzergenerierten Inhalte an das Modell stellen, liefert es für die deutsche Sprache gute Ergebnisse.
Beispielsweise wird die Bewertung

„super schick ! kann ich nur empfehlen ! Gute Verarbeitung“

korrekt unterteilt in

„super schick !“, „kann ich nur empfehlen !“, „Gute Verarbeitung“

Jedoch kann die Bewertung

„Leider viell zu kurz.Schade.Ansonsten hüpsch.Musste leider zurück“

nicht korrekt getrennt werden, da jeweils das Leerzeichen nach dem Punkt fehlt.

2.6.4 Tokenisierung

Tokenisierung bezeichnet die Einteilung von Texten in einzelne linguistische Einheiten [Car+09, Kap. 3.4]. Eine linguistische Einheit ist zum Beispiel ein Wort oder ein Satzzeichen.
Nachdem die Bewertungen in Sätze segmentiert wurden, wird jeder Satz in Worte unterteilt.

Sowohl für das Part-of-Speech Tagging (siehe Kapitel 3.1), als auch für das Training der Word2Vec Modelle (siehe Kapitel 4.1) müssen die Bewertungen nicht nur in Sätze, sondern auch in Worte unterteilt werden.
Zur Tokenisierung der Sätze wurde *enchant* (siehe Kapitel 2.3) verwendet. Dieses bietet nur einen Tokenizer für die englische Sprache, welcher aber auch im Deutschen gut funktioniert.
Wie in Kapitel 2.6.3 gezeigt werden Sätze, bei denen das Leerzeichen nach dem Punkt fehlt, nicht korrekt getrennt. Dies führt dazu, dass die Worte vor und nach dem Punkt bei der Tokenisierung als ein Wort erkannt werden. Um diese Fehler zu beheben wurden die durch die Tokenisierung gefundenen Worte zusätzlich getrennt, wenn sie einen Punkt enthalten.

2.7 Wikipedia Daten

Zusätzlich zu den Bewertungsdaten wurden die Daten von der deutschen Wikipedia heruntergeladen[14]. Diese wurden bei der Merkmalsextraktion mit Word2Vec als zusätzliche Trainingsdaten verwendet (siehe Kapitel 4.2.3).
Wikipedia wurde als zusätzliche Datenbasis genutzt, da die deutsche Wikipedia nicht nur eine große, sondern auch eine leicht zugänglichen Datenbasis für deutsche Texte darstellt.
Die deutsche Wikipedia enthält rund $72,5$ Millionen Sätze und $364,9$ Millionen Worte.
Um die Daten für das Training von Word2Vec Modellen aufzubereiten wurden sie zunächst bereinigt.
Das heißt es wurden alle Links, Referenzen, XML-Tags und ähnliches entfernt. Dies wurde mit einem Perl Script von Matt Mahoney[15] (auch im Anhang A.2) erreicht, welches auch von Mikolov et al. auf der Google Code Seite von Word2Vec vorgeschlagen wird[16]. Dieses Script ist speziell auf die Bereinigung von Wikipedia Daten zugeschnitten und wurde lediglich um eine Zeile für das Filtern der Zeichen „Ä, Ö, Ü, ä, ö, ü, ß“ erweitert, sowie um die Ersetzung der englischen Zahlworte reduziert, um es für die deutsche Sprache anzupassen.

[14] http://dumps.wikimedia.org/dewiktionary/20141124/, abgerufen am 29.12.2014
[15] http://mattmahoney.net/dc/textdata.html, abgerufen am 15.01.2015
[16] https://code.google.com/p/word2vec/, abgerufen am 26.01.2015

Kapitel 3

Merkmalsextraktion durch häufigste Nomen

In diesem Kapitel wird die Methode der Merkmalsextraktion evaluiert, welche darauf basiert die häufigsten Nomen der Bewertungen herauszufiltern. Dabei werden zunächst die Bewertungen mit Hilfe des Part-of-Speech Taggings mit Wortarten (siehe Anhang A.1) annotiert und danach die häufigsten Nomen und Eigenworte betrachtet.
Dieser Ansatz wurde, wie in Kapitel 1.4 beschrieben, schon mehrfach verwendet und auch in [Ton11] im Speziellen für Bewertungen in deutscher Sprache angewendet.
Der Vorteil dieses Ansatzes ist es, dass das Part-of-Speech Tagging ein gut untersuchtes Gebiet ist und mittlerweile auch für Webinhalte in deutscher Sprache zuverlässig gute Ergebnisse liefert [GE09].
In den Part-of-Speech getaggten Daten die Nomen zu filtern ist eine triviale Aufgabe. Die Hauptaufgabe besteht darin aus den häufigsten Nomen die tatsächlichen Merkmale zu extrahieren. Dazu wird auch das Wissen aus der Domäne genutzt, indem Marken und Kategorienamen gefiltert werden, da diese keine Merkmale darstellen.

3.1 Part-of-Speech Tagging

Zunächst müssen die Bewertungen mit Wortarten annotiert werden. Als Part-of-Speech (POS) Tagging bezeichnet man den Vorgang, in welchem den Worten und Satzzeichen des zu analysierenden Textes Wortarten zugeordnet werden.
Für die deutsche Sprache steht hierzu das Stuttgart-Tübingen Tagset oder auch kurz STTS[1] zur Verfügung (siehe auch Anhang A.1).
Das Part-of-Speech Tagging wird mit Hilfe von *NLTK* realisiert. Es wurde der Stanford Tagger verwendet, welcher von *NLTK* mitgeliefert wird. Im Speziellen wurde daraus das dewac Modell genutzt. Der Begriff *dewac* steht in diesem Fall für *deutsches Web als Korpus*. Dieses Modell wurde mit dem Negra Korpus [Kru+11] trainiert, wobei zusätzlich die Merkmale aus den Clustern ähnlicher Wortverteilungen (distributional similarity [PTL93]) verwendet werden, welche aus dem wac (web as corpus) [BK06] Korpus gewonnen wurden. Dieser Korpus besteht aus mehr als einer Milliarde Tokens, die Texten von deutschen Webseiten entnommen wurden [Car+09, Kap. 4.1]. Die Verbesserung des Taggers mit Hilfe des dewac Korpus sorgt dafür, dass dieser Tagger besser

[1] http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html

für Daten aus dem Onlinebereich geeignet ist.
Der Stanford Tagger braucht im Vergleich zu anderen Taggern sehr lange um die Worte zu annotieren, hat dafür aber eine höhere Genauigkeit bei der POS Annotation von Worten, insbesondere bei unbekannten Worten [GE09]. Bei Giesbrecht und Evert wurden die Tagger dabei auf dem deutschen TIGER Korpus[2] trainiert und mit einem Teil des *dewac* Korpus getestet. Der Trigrams'n'Tags (TnT) [Bra00] Tagger, welcher auf dem Hidden Markov Model basiert, hat mit 92,69% richtig annotierten Worten eine leicht höhere gesamte Genauigkeit bei der Annotation als der Stanford Tagger mit 92,61%. Jedoch ist der Stanford Tagger mit 75,35% richtig POS annotierten Worten deutlich besser als der TnT Tagger mit 71,99% [GE09].
Bei Giesbrecht und Evert wurde eine ältere Version des Stanford Taggers verwendet, welche noch nicht mit den Merkmalen aus dem dewac Korpus verbessert wurde. Daher und wegen der höheren Genauigkeit bei der Annotation von unbekannten Worten wurde dieser aktuellere Tagger gegenüber dem TnT Tagger bevorzugt um die Daten aus dem Onlinebereich zu annotieren.

3.2 Häufigste Nomen

Für die Extraktion der häufigsten Nomen wurden 455351 Bewertungen von 1762 Produkten aus dem Bereich Damenmode Part-of-Speech getaggt.
Hier wurde nur ein Teil der Daten aus dem Bereich Damenmode getaggt, da das Part-of-Speech Tagging zu lange braucht um alle Bewertungen zu annotieren (siehe Kapitel 3.3).
Von den POS getaggten Bewertungen wurden nach manueller Durchsicht maximal 0.5% der häufigsten Nomen und Eigenworte als Kandidaten für Produktmerkmale in Betracht gezogen. Wie in [Ton11] vorgeschlagen, wurden nicht nur die häufigsten Nomen, sondern auch die häufigsten als Eigennamen getaggten Worte in die Liste mit aufgenommen.

3.2.1 Filterung

Viele der häufigsten Nomen und Eigennamen sind Marken oder Kategorienamen. Diese stellen keine Produktmerkmale dar und werden daher herausgefiltert. Wie die Kategorienamen aus dem Kategoriebaum von *otto.de* können auch die Marken automatisiert aus der Seite ausgelesen werden, da sie auf der Startseite der Unterkategorie Damenmode aufgelistet werden.
Da die Texte oft unvollständige Sätze oder nur Stichworte enthalten, kommt es beim Part-of-Speech Tagging häufig zu Fehlern und Worte werden fälschlicherweise als Nomen oder Eigennamen getaggt. Einige dieser falsch getaggten Worte sind Stopworte. Daher wurden zusätzlich auch die Stopwörter herausgefiltert.
Da für das POS-Tagging vollständige Sätze, inklusive Stopworten, benötigt werden, werden diese erst jetzt herausgefiltert.
Stopworte sind Worte, welche sehr oft in Dokumenten auftauchen, aber keine Aussage über den Inhalt eines Dokumentes erlauben. Beispiele für deutsche Stopworte sind „ein“, „der“, oder „auch“.

[2] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.en.html, abgerufen am 27.01.2015

In dieser Arbeit wurde eine Liste deutscher Stopworte[3] aus einer Sammlung von Stopworten in verschiedenen Sprachen verwendet[4]. Diese wurde um weitere Worte ergänzt (wie *ca* und *daraufhin*) und Worte, welche als Meinungen in Frage kommen (wie *schlecht* und *großer*), wurden aus der Liste gelöscht.
Die Python Bibliothek *NLTK* bietet ebenfalls eine Liste deutscher Stopworte. Die verwendete Liste ist jedoch deutlich umfangreicher und musste nicht mehr stark erweitert werden.

3.2.2 Vergleich verschiedener Anteile an häufigsten Nomen und Eigennamen

Wie bereits erwähnt, wurden zusätzlich zu den Nomen auch die häufigsten Eigennamen betrachtet, da unbekannte Worte beim Part-of-Speech Tagging oft als Eigennamen annotiert werden.
Die Ergebnisse der Extraktion der Nomen und Eigennamen aus den Part-of-Speech getaggten Bewertungen sind in Tabelle 3.1 dargestellt. Diese gibt auch einen Überblick darüber wie viele der gefundenen Worte tatsächlich Merkmale sind, wenn nur als Nomen oder nur als Eigennamen annotierte Worte oder eine Kombination aus beiden betrachtet wird.
Nach manueller Durchsicht der häufigsten Nomen und Eigennamen wurden maximal die häufigsten 0,5% als Merkmale in Betracht gezogen. Wie in Tabelle 3.1 gezeigt sind auch bei einer relativ kleinen Menge von nur 0,5% der häufigsten Nomen auch nach der Filterung die meisten Worte keine tatsächlichen Merkmale. Von den als Eigenworten annotierten Worten sind nur 7,5% der gefilterten Worte tatsächlich Merkmale. Von den Nomen sind nach der Filterung immerhin noch fast 30% tatsächlich Merkmale. Eine Liste dieser Worte befindet sich im Anhang B.1. Anzumerken ist, dass Nomen aus der Liste der häufigsten Nomen, welche keine Merkmale darstellen, meist Körperteile (Oberweite, Brust, Busen, Beine, Hals, ...) oder Personen (Frau, Frauen, Mann, Tochter, ...) bezeichnen. In Tabelle 3.2 ist dargestellt wie sich das Verhältnis von gefilterten häufigsten Nomen zu tatsächlichen Merkmalen verhält, wenn weniger als die bisher angenommenen 0,5% der häufigsten Nomen betrachtet werden. Auch durch die Verringerung des Anteils der betrachteten häufigsten Nomen können keine signifikant besseren Resultate erzielt werden. Verringert man den Anteil der betrachteten häufigsten Nomen so sind zwar mehr gefilterte Worte tatsächlich Merkmale, jedoch sinkt die absolute Zahl der tatsächlichen Merkmale stärker. Dies führt dazu, dass bei ei-

Tabelle 3.1: 0.5% häufigste Nomen und Eigennamen

Tags	ohne Filter	mit Filter	Nomen	Nomen %	Merkmale	Merkmale %
NN + NE	330	283	213	75,3%	68	24,0%
NN	278	234	194	82,9%	66	28,2%
NE	91	80	20	25,0%	6	7,5%

[3] http://members.unine.ch/jacques.savoy/clef/germanST.txt
[4] http://members.unine.ch/jacques.savoy/clef/, abgerufen am 13.01.2015

Tabelle 3.2: Vergleich verschiedener Anteile der häufigsten Nomen

Anteil häufigste Nomen	ohne Filter	mit Filter	Merkmale	Merkmale %
0,5%	278	234	66	28,2%
0,3%	158	119	41	34,5%
0,1%	52	37	18	48,6%

nem Anteil der 0,1% häufigsten Nomen nach der Filterung annähernd 50% tatsächlich Merkmale sind, jedoch sind dies absolut nur noch 18 Worte.

3.3 Evaluation der Merkmalsextraktion durch häufigste Nomen

Die Schwierigkeit mittels der Häufigkeit der Nomen Merkmale zu finden besteht darin, dass hier nicht zwischen dem Produkt selbst, einer Person oder einem Merkmal unterschieden werden kann, da diese alle als Nomen getaggt werden.
In anderen Arbeiten wurden für das Auffinden von Merkmalen auch die Worte in direkter Nähe zu den gefundenen Nomen, sowie ihr Part-of-Speech Tag betrachtet. Bei [HL04] beispielsweise werden seltene Merkmale gefunden, indem in der Nähe von Meinungsworten nach Nomen gesucht wird. Da jedoch die Ergebnisse der häufigsten Nomen zeigen, dass beim Part-of-Speech Tagging sehr viele fehlerhafte Annotationen erzeugt wurden, wurde dieser Ansatz nicht weiter verfolgt.
Ein weiterer Nachteil dieser Methode der Merkmalsextraktion besteht darin, dass das Part-of-Speech Tagging sehr zeitaufwändig ist. Der gesamte Prozess der Vorverarbeitung inklusive der Rechtschreibkorrektur und dem Part-of-Speech Tagging hat mehr als acht Wochen in Anspruch genommen[5].
Es zeigt sich, dass mit dieser Methode nur verwertbare Ergebnisse erzeugt werden können, wenn vor allem die Vorverarbeitungsschritte der Satzsegmentierung und der Rechtschreibkorrektur gute Ergebnisse liefern. Sind die Ergebnisse dagegen wie hier nur von geringer Qualität, so können auch beim Part-of-Speech Tagging die Texte nicht zuverlässig POS annotiert werden.
Aus den hier generierten Resultaten wird geschlossen, dass mit den vorliegenden Daten ohne eine grundlegende Änderung und Verbesserung der vorhergehenden Prozessschritte Rechtschreibkorrektur und Satzsegmentierung keine Ergebnisse erzielt werden können, mit denen eine zuverlässige Merkmalsextraktion möglich ist.

[5] Das Programm lief dabei auf einer Amazon EC2 Spot Instanz mit 16 Kernen mit je 2.50 GHz und 128 GB Ram.

Kapitel 4

Merkmalsextraktion mit Word2Vec

Nachfolgend wird erläutert was Word2Vec ist und welche Parameter und Algorithmen bei der Berechnung von Word2Vec Modellen zur Verfügung stehen. Außerdem werden einige der besten Modelle im Detail vorgestellt und deren Genauigkeit im Bezug auf die Beantwortung von semantischen und syntaktischen Fragen verglichen. Auf die Erstellung dieser Fragen wird näher eingegangen.
Anschließend wird eine Methode vorgestellt, mit welcher mittels eines Word2Vec Modells Merkmale in Bewertungstexten gefunden werden können, und diese anhand von Metriken evaluiert.
Die Modelle, welche mit den Bewertungen trainiert wurden, enthalten jeweils die Bewertungen von allen Kategorien Damenmode, Herrenmode und Multimedia.

4.1 Word2Vec

Mit Word2Vec werden Worte als Vektoren repräsentiert. Wörtlich heißt Word2Vec Wort zu Vektor. Dabei wird mittels *Distributed representation*[Mik+13b] ein n-dimensionaler Vektorraum erzeugt, bei dem jedes Wort aus den Trainingsdaten durch einen Vektor repräsentiert wird.
Im nächsten Schritt werden die Vektoren in ein neuronales Netz eingespeist, welches mit Hilfe eines Lernalgorithmus die Vektoren so verändert, dass ähnliche Worte einen ähnlichen Vektor bilden.
Die Ähnlichkeit zwischen den Vektoren wird mit Hilfe der Kosinus-Distanz berechnet (siehe Kapitel 4.1.6).
Zur Berechnung der Wortvektoren können zwei verschiedene Architekturen von neuronalen Netzen genutzt werden, Continuous bag-of-words (CBOW) und Skip-gram. Um diese neuronalen Netze zu trainieren stehen zwei verschiedene Lernalgorithmen zur Auswahl. Dies sind zum einen *hierarchical softmax* und zum anderen *negative sampling*. Für das Training können verschiedene Parameter eingestellt werden.

4.1.1 Parameter

Im Folgenden werden die einzelnen Parameter für das Lernen der Wortvektoren mit Word2Vec kurz vorgestellt.

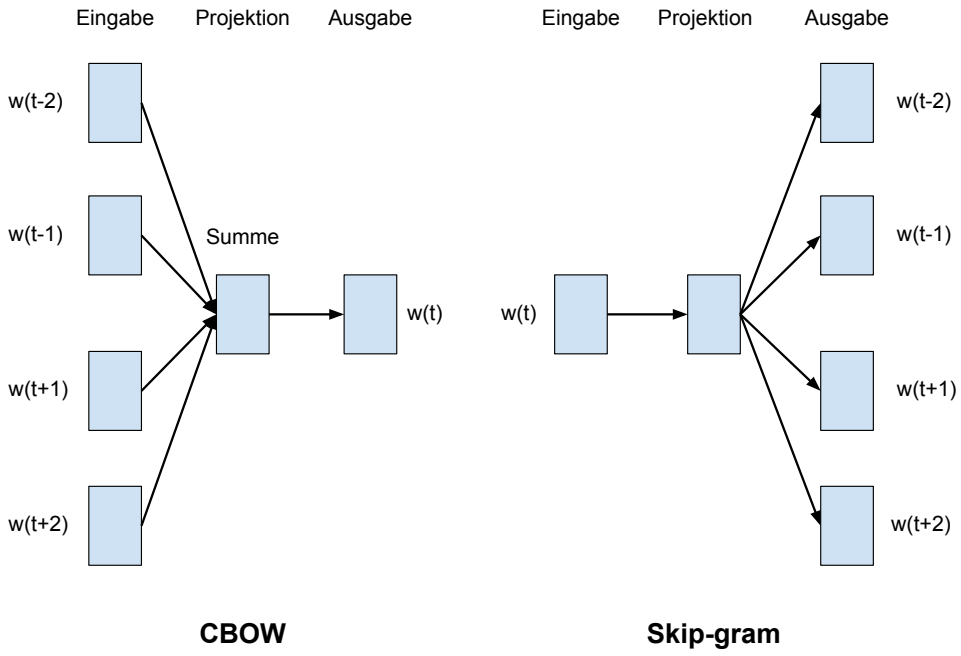

Abbildung 4.1: CBOW und Skip-gram im Vergleich, nach [Mik+13b], übersetzt

size

Die Anzahl der Dimensionen der Wortvektoren wird mit dem Parameter *size* eingestellt. Wird also von einem n-dimensionalen Vektorraum gesprochen, so steht n in diesem Fall für den Wert des Parameters *size*.

window

window ist die Anzahl der benachbarten Worte, welche für die Berechnung der Wortvektoren mit betrachtet werden.

min_count

Der Wert des Parameters *min_count* gibt an wie oft ein Wort im Korpus vorkommen muss um ins Wörterbuch aufgenommen zu werden. Je größer der Trainingskorpus ist, mit welchem das Wörterbuch gebildet wird, desto höher kann dieser Wert eingestellt werden.

negative

Der Parameter *negative* wird nur benötigt, wenn als Lernalgorithmus das negative sampling verwendet wird. Für eine Erklärung zum Negative sampling siehe Kapitel 4.1.4.

4.1.2 CBOW

Die Abkürzung CBOW steht für Continuous bag-of-words (zu Deutsch stetige Menge von Worten). Es ist ein neuronales Netz ohne verdeckte Schicht (hidden layer). Dabei wird ein Wort anhand des Kontextes vorhergesagt (siehe Abbildung 4.1). Die Menge der Worte, die dabei als Kontext verwendet werden, wird über den Parameter *window* eingestellt.

4.1.3 Skip-gram

Mittels des Skip-gram Algorithmus wird anders als beim CBOW Algorithmus aus einem gegebenen Wort der Kontext vorhergesagt (siehe Abbildung 4.1). Die Menge der vorhergesagten Worte wird über den Parameter *window* definiert. Wie auch beim CBOW-Algorithmus handelt es sich hierbei um ein neuronales Netz ohne verdeckte Schicht (hidden layer). Weniger komplexe Modelle wurden bei der Entwicklung von Word2Vec gegenüber tiefen neuronalen Netzen mit verdeckten Schichten bevorzugt, da diese schneller und daher effizienter mit mehr Daten trainiert werden können[Mik+13b]. Für die weitere Berechnung der Word2Vec Modelle wurde der Skip-gram Algorithmus gewählt, da bei [Mik+13b] hiermit die besten Ergebnisse erzielt wurden und der Algorithmus daher auch in [Mik+13a] noch erweitert und verbessert wurde.
Eine weitere Begründung, warum der Skip-gram gegenüber dem CBOW Algorithmus bevorzugt wurde, findet sich in Tabelle 4.1. Der Skip-gram Algorithmus lieferte in Verbindung mit dem hierarchical softmax (in der Tabelle Spalte *hs*) Lernalgorithmus das beste Ergebnis bezogen auf die Anzahl der richtig beantworteten Fragen aus dem Fragenkatalog (siehe Kapitel 4.2).

4.1.4 Negative sampling

Das negative sampling kann alternativ zum hierarchical softmax als Lernalgorithmus für das neuronale Netz genutzt werden [Mik+13a].
Bei diesem Ansatz wird für ein gegebenes Wort nicht die Ähnlichkeit zu allen anderen Worte berechnet, sondern es wird andersherum davon ausgegangen, dass zufällig gewählte Worte dem gegebenen Wort mit hoher Wahrscheinlichkeit unähnlich sind. Die Anzahl dieser zufällig gewählten *negative samples* kann mit dem Parameter *negative* angegeben werden.
In Tabelle 4.1 zeigt sich, dass bei den hier genutzten Trainingsdaten der negative sampling Lernalgorithmus schlechtere Ergebnisse liefert als der hierarchical softmax Algorithmus. Es wurde der negative sampling Lernalogrithmus jeweils in Verbindung mit dem CBOW Netz, als auch dem Skip-gram Netz verwendet. Jeweils einmal mit einem Wert für den Parameter *negativ* von 5 (Spalte negS-5) und einmal mit *negative*=15 (Spalte negS-15). Bei [Mik+13a] wird für Parameter *negative* ein Bereich von 2-20 als gut angegeben.

4.1.5 Hierarchical softmax

Hierarchical softmax ist eine effizientere Alternative zum herkömmlichen softmax Lernalgorithmus [Mik+13a].
Nutzt man hierarchical softmax als Lernalgorithmus so wird das Wörterbuch als Huffman Binärbaum aufgebaut. Dabei haben die häufigsten Worte die kürzeste Codierung

[Huf52]. Dies führt dazu, dass weniger Speicherplatz benötigt wird und auch die Berechnungen schneller ablaufen [Mik+13b]. Die Berechnung der Vektoren erfolgt in zwei

Tabelle 4.1: Vergleich von CBOW und Skip-gram, sowie hierarchical softmax und negative sampling

	hs	negS-5	negS-15
Skip-gram	14,9%	12,8%	12,6%
CBOW	12,0%	11,1%	10,4%

Schritten.
Im ersten Schritt wird aus den verwendeten Trainingsdaten ein Wörterbuch erstellt. Hierbei können Stopwörter entfernt werden. Ebenso können mit Hilfe des Parameters *min_count* Worte entfernt werden, welche selten im Korpus vorkommen, da für diese keine aussagekräftigen Vektoren berechnet werden können.
Im zweiten Schritt wird für jedes Wort im Korpus mittels *Distributed representation*[Mik+13b] ein n-dimensionaler (n=*size*) Wortvektor berechnet.
Wie in Tabelle 4.1 zu sehen ergibt die Kombination der Skip-gram Architektur zusammen mit dem hierarchical softmax Lernalgorithmus das beste Ergebnis (zur Erklärung der Werte siehe nächstes Kapitel 4.2). Daher wurden alle folgenden Word2Vec Modelle mit dieser Kombination aus Skip-gram neuronalem Netz und hierarchical softmax Lernalgorithmus trainiert.
Bei der Berechnung eines Word2Vec Modells aus allen Bewertungen (ca. 1600000 Sätze) mit 600 Dimensionen benötigt die Berechnung des Wörterbuches und der initialen Wortvektoren rund 13 Minuten. Das Training mit allen Bewertungen dauert rund 17 Minuten[1]. Die Zeit zur Berechnung steigt annähernd linear mit der Menge an Trainingsdaten.

4.1.6 Distanz zwischen Vektoren in Word2Vec

Die Distanz zwischen den einzelnen Vektoren wird in Word2Vec mit Hilfe der Kosinus-Ähnlichkeit berechnet.

$$cos(\vec{X}, \vec{Y}) = \frac{\Sigma_{i=1}^{n} x_i y_i}{\sqrt{\Sigma_{i=1}^{n} x_i^2} \sqrt{\Sigma_{i=1}^{n} y_i^2}} \tag{4.1}$$

Die Kosinus-Ähnlichkeit berechnet sich aus dem Kosinus des Winkels zwischen zwei Vektoren (siehe Formel 4.1). Dabei steht n für die Anzahl der Dimensionen der Vektoren X und Y. Der Wertebereich für die Kosinus-Ähnlichkeit liegt bei -1 bis +1, wobei in diesem Fall nur Werte zwischen 0 und 1 erreicht werden, da der Winkel zwischen den Wortvektoren maximal 90° beträgt.

4.2 Evaluation der Word2Vec Modelle

Um die Genauigkeit der gelernten Word2Vec Modelle im Bezug auf das Erkennen von semantischen und syntaktischen Beziehung zwischen Worten zu testen, wurde nach dem

[1] Eingesetzte Hardware: MacBook Pro mit 8GB Ram und 2 Kernen mit je 2,3GHz.

Beispiel von [Mik+13b] eine Liste semantischer und syntaktischer Fragen erstellt[2].
Dieser Ansatz zur Bewertung der Qualität der Word2Vec Modelle beruht auf der Tatsache, dass mit einfachen Vektorrechnungen die semantischen und syntaktischen Beziehungen zwischen Worten abgebildet werden können. Das bekannteste Beispiel ist die Formel 4.2.

$$\vec{woman} + \vec{king} - \vec{man} = \vec{queen} \tag{4.2}$$

Hierbei werden die Vektoren für *Frau* und *König* addiert und der Vektor für *Mann* subtrahiert. Der daraus resultierende Vektor ist dem Vektor für *Königin* am nächsten. Nach diesem Muster lassen sich nicht nur wie hier semantische Zusammenhänge, sondern auch syntaktische Zusammenhänge aufzeigen. In Formel 4.3 wird beispielsweise der Zusammenhang zwischen einem Adjektiv und dessen Komparativ abgebildet.

$$\vec{gut} + \vec{schlechter} - \vec{schlecht} = \vec{besser} \tag{4.3}$$

Bei [Mik+13b] wurde die Liste der semantischen und syntaktischen Fragen erstellt, indem manuell Wortpaare zu verschiedenen Kategorien gesammelt und diese dann zu Fragen kombiniert wurden, nach dem Schema *a zu b verhält sich wie c zu ?*.
Die dort verwendeten Kategorien umfassen unter anderem Gegenteile, Singular und Plural von Nomen und Verben, sowie Frau zu Mann Beziehungen.
Da die Fragen in [Mik+13b] für die englische Sprache erstellt wurden und auch die Daten aus einem anderen Kontext kommen, wurde ein komplett neues Set an syntaktischen und semantischen Fragen erstellt.
Teilweise wurden die Wortpaare aus dem Original übersetzt und teilweise mit Wortpaaren aus dem Kontext Mode und Bewertungen ergänzt.
Insgesamt besteht das Set aus 3360 semantischen und syntaktischen Fragen, (siehe B.4) jeweils gleichmäßig verteilt auf die 8 Bereiche:

- Gegenteil
- Perfekt erste Person Singular
- Präteritum dritte Person Singular
- Steigerung
- Plural Nomen
- Geschlecht
- Superlativ
- Präteritum erste Person Singular

Hierbei bildet nur der Bereich *Geschlecht* einen semantischen Zusammenhang ab. Die weiteren in [Mik+13b] erwähnten Bereiche zu semantischen Zusammenhängen wie zum Beispiel Hauptstadt zu Land Beziehungen sind für die Bewertungsdaten irrelevant und werden daher nicht betrachtet.
Das Set der semantischen und syntaktischen Fragen wird im Folgenden als Fragenkatalog bezeichnet.
Der Aufbau der Datei erfolgt nach dem folgenden Schema:

- Die erste Zeile eines neuen Bereiches beginnt mit einem Doppelpunkt und nachfolgend der Bezeichnung des Bereiches (zum Beispiel *: Gegenteil*).

[2] https://code.google.com/p/word2vec/source/browse/trunk/questions-words.txt, abgerufen am 28.12.2014

- Dieser ersten Zeile folgen in jeder Zeile 4 Worte, welche zusammen die semantische oder syntaktische Frage nach dem Schema *a verhält sich zu b wie c zu d*, wobei das Wort d vom Modell vorhergesagt werden soll.

Wird also nachfolgend von der Accuracy oder Genauigkeit der Modelle gesprochen, so bezieht sich das auf die Anzahl der Worte d, welche vom Modell korrekt vorhergesagt wurden.

4.2.1 Vergleich der besten Modelle

Es wurden mehrere Modelle mit verschiedenen Parametern und Trainingsdaten erstellt. Die Modelle, für welche die höchste Genauigkeit bei der Beantwortung der semantischen und syntaktischen Fragen erreicht wurde, werden im folgenden Abschnitt beschrieben und verglichen. Eine Übersicht der betrachteten Modelle ist in Tabelle 4.2 dargestellt. Die Accuracy (Genauigkeit) ist hierbei die Anzahl der richtig beantworteten Fragen aus dem Fragenkatalog. Die Ergebnisse sind dabei wie folgt aufgeschlüsselt:
Für jeden Bereich des Fragenkataloges wird absolut und als Anteil aufgezeigt auf wie viele Fragen das Modell das erwartete Ergebnis liefert. Der Anteil bezieht sich dabei nicht auf die gesamte Anzahl der Fragen im jeweiligen Bereich, sondern nimmt Bezug auf den Teil der Fragen, bei denen alle Worte im Modell vorkommen. Dieser Teil, sowie die absolute Anzahl der richtig beantworteten Fragen, steht in der Spalte *Accuracy absolut*.
Die Daten, die zur Erstellung der Modelle genutzt wurden, wurden nicht mittels der Rechtschreibkorrektur korrigiert. Mit korrigierten Daten wurden deutlich schlechtere Ergebnisse erzielt. Bei den Modellen, welche mit den korrigierten Daten aus den Bewertungen trainiert wurden, ist die Accuracy im Durchschnitt 1,7% schlechter als bei den Modellen, bei denen keine Rechtschreibkorrektur verwendet wurde.
Warum die Rechtschreibkorrektur an sich so schlechte Ergebnisse liefert ist in Kapitel 2.6.2 dargestellt.
Um die Menge der Worte der Wörterbücher für die Modelle zu verringern und zu verhindern, dass ein Wort in verschiedenen Versionen der Groß-/Kleinschreibung in den Modellen existiert, wurden bei der Berechnung der Word2Vec Modelle alle Buchstaben in den Worten auf Kleinbuchstaben zurückgeführt.

Tabelle 4.2: Modelle in der Übersicht

Modell	Kürzel	size	Bi-gramme	min_ count	Anzahl Worte	Ab-deckung	Accuracy
Bewertungen	Reviews	120	Ja	10	10,9 Mio	69%	16,0%
Wikipedia	Wiki	300	Nein	5	364,9 Mio	50%	32,0%
Wikipedia + Bewertungen	Wiki_1	300	Nein	5	375,8 Mio	77%	28,5%
Wikipedia + Bewertungen 2	Wiki_2	600	Nein	5	375,8 Mio	77%	30,5%
Wikipedia + Bewertungen 3	Wiki_3	600	Nein	10	375,8 Mio	69%	30,9%

4.2.2 Modell mit Trainingsdaten aus den Bewertungen (*Reviews*)

Das erste Modell wurde nur mit den Daten aus allen gesammelten Bewertungen (siehe Tabelle 2.2) erstellt. Es wurden verschiedene Parameter getestet und die Werte für die so erhaltenen Genauigkeiten verglichen.
Es wurden nur Worte in das Wörterbuch des Modells aufgenommen, welche minimal 10 mal im Trainingskorpus auftreten. Dadurch ergibt sich eine Abdeckung der Worte aus dem Fragenkatalog von 69%. Abdeckung bedeutet in diesem Fall, dass 69% der Worte aus dem Fragenkatalog tatsächlich auch im Wörterbuch des Modells enthalten sind.
Die Anzahl der Dimensionen der Wortvektoren wurde nach einigen Tests mit unterschiedlichen Werten auf *size*=120 festgelegt.
Es werden jeweils die 8 nächsten Worte (*window*=8) zur Berechnung der Vektoren betrachtet. Dies hat sich als ein Wert erwiesen, welcher die besten Ergebnisse liefert und wird daher auch in den folgenden Modellen verwendet. Bei Tests mit *window*=5 sank die Genauigkeit um 0,7%, bei *window*=10 um 1,6%.
Für die Berechnung der Wortvektoren werden die Daten Satz für Satz eingelesen. Da es vorkommt, dass Sätze nicht korrekt getrennt werden, ist es sinnvoll den Parameter *window* auf einen Wert von 8 zu begrenzen. Dies verhindert, dass Worte über die eigentliche Satzgrenze hinweg zur Berechnung der Vektoren mit einbezogen werden.
Vor der Erstellung des Wörterbuches wurden Bigramme erstellt. Dabei werden Worte, welche häufig zusammen auftreten, zu einer Phrase zusammengefasst. Dies ist mit *gensim* mittels des Moduls *Phrases* möglich.
Tabelle 4.3 zeigt die Ergebnisse der Genauigkeit des beschriebenen Modells im Bezug auf die Beantwortung der semantischen und syntaktischen Fragen. Die gesamte Accuracy, also der Anteil der richtig beantworteten Fragen aus dem Fragenkatalog, beträgt 16,0%. Dies ist ein relativ geringer Wert, welcher der kleinen Zahl an Trainingsdaten geschuldet ist. Um die semantischen und syntaktischen Zusammenhänge deutlich abbilden zu können, müssen die Worte in ausreichender Anzahl in den Trainingsdaten vorkommen.

Tabelle 4.3: Accuracy Modell *Reviews*

Kategorie	Accuracy	
	[%]	absolut
Gegenteil	9,9	(27/272)
Perfekt erste Person Singular	9,9	(27/272)
Präteritum dritte Person Singular	24,9	(85/342)
Steigerung	21,0	(88/420)
Plural Nomen	12,1	(51/420)
Geschlecht	28,2	(59/209)
Superlativ	6,7	(2/30)
Präteritum erste Person Singular	8,8	(30/342)
total	16,0	(369/2307)

Dies fällt besonders im Bereich *Superlative* auf, in dem die Accuracy nur 6,7% beträgt. In diesem Bereich sind nur von 30 der insgesamt 420 Fragen alle Worte im Wörterbuch vorhanden sind, was darauf schließen lässt, dass allgemein in den Bewertungen seltener Superlative verwendet werden. Dadurch ist der syntaktische Zusammenhang in den

Vektoren nicht so deutlich.
Besonders in den Bereichen *Geschlecht*, *Präteritum dritte Person Singular* und *Steigerung* ist die Accuracy im Vergleich zur gesamten Accuracy überdurchschnittlich gut.
Dabei ist zu bemerken, dass im Bereich *Geschlecht* nur etwa bei der Hälfte der Fragen alle Worte im Modell vorkommen. Die Worte, die jedoch im Modell vorkommen, lassen einen starken semantischen Zusammenhang erkennen.
In den Bereichen *Präteritum dritte Person Singular* und *Steigerung* ist jeweils die Abdeckung mit 81,4% und 100% sehr hoch. Dies lässt darauf schließen, dass die Worte in diesem Bereich des Fragenkataloges oft in den Trainingsdaten vorkommen, was den syntaktischen Zusammenhang verstärkt.

4.2.3 Modell mit Trainingsdaten aus Wikipedia (*Wiki*)

Um zu sehen wie sich die Accuracy (Genauigkeit) bei der Beantwortung der semantischen und syntaktischen Fragen verhält, wenn die Vektoren mit einem redaktionell erstellten Set von Trainingsdaten berechnet werden, wurde ein Modell erstellt, welches mit dem kompletten Auszug der deutschen Wikipedia (siehe Kapitel 2.7) trainiert wurde.
Für die Berechnung dieses Modells wurde der Parameter *size*, also die Dimension der Wortvektoren, auf einen Wert von 300 erhöht. Die Ergebnisse aus [Mik+13b] zeigen, dass eine höhere Dimension von Wortvektoren, vor allem in Verbindung mit mehr Trainingsdaten, zu einem deutlich besseren Ergebnis führt.
Um die Abdeckung zu erhöhen wurde die minimale Häufigkeit eines Wortes für die Aufnahme in das Wörterbuch auf einen Wert von 5 herabgesetzt.
Auch wurden hier keine Bigramme berechnet, da bei dieser Menge an Daten die Berechnung sehr lange dauert. Bei der Berechnung der Modelle mit den Bewertungsdaten allein steigert sich die Berechnungsdauer um 40%, wenn Bigramme verwendet werden.

In Tabelle 4.4 wird ersichtlich, dass deutlich weniger Worte aus dem Fragenkatalog überhaupt, oder häufig genug, in den Daten vorkommen. Die Abdeckung mit den Daten aus Wikipedia liegt also, trotz der Reduktion der minimalen Häufigkeit, deutlich unter der des Modells aus den Bewertungsdaten. Dies liegt daran, dass der Fragenkatalog für Daten aus dem Bereich Mode erstellt wurde, insbesondere für Bewertungen. Vor allem Verbformen zur ersten oder dritten Person Singular kommen deutlich seltener vor als bei den Bewertungsdaten.
Insgesamt beträgt die Abdeckung der Worte aus dem Fragenkatalog nur etwas mehr als 50%. Jedoch wird fast ein Drittel der Fragen, deren Worte im Wörterbuch vorkommen, richtig beantwortet.
Auffällig ist, dass hier weniger als die Hälfte der Superlative im Wörterbuch enthalten sind. Auch werden hier, wie bei den Bewertungsdaten, die Testfragen zu den Superlativen nur im einstelligen Prozentbereich richtig beantwortet.

Tabelle 4.4: Accuracy Modell *Wiki*

Kategorie	Accuracy	
	[%]	absolut
Gegenteil	30,8	(74/240)
Perfekt erste Person Singular	12,9	(17/132)
Präteritum dritte Person Singular	64,3	(117/182)
Steigerung	35,1	(120/342)
Plural Nomen	13,7	(25/182)
Geschlecht	55,1	(168/305)
Superlativ	2,6	(4/156)
Präteritum erste Person Singular	13,7	(25/182)
total	32,0	(550/1721)

Neben dem Bereich *Superlative* ist die Accuracy in den Bereichen *Perfekt erste Person Singular* und *Präteritum erste Person Singular* auffallend niedrig. Wie erwähnt ist dort auch die Abdeckung deutlich geringer als beim Modell aus den Bewertungsdaten. Die geringe Accuracy ist der Art der Daten geschuldet. Da Wikipedia ein Onlinelexikon zur Sammlung von Wissen ist[3], wird hier selten aus der Ich-Perspektive (erste Person Singular) geschrieben. Dadurch ergibt sich nur ein schwacher syntaktischer Zusammenhang zu den Bereichen *Perfekt erste Person Singular* und *Präteritum erste Person Singular* in den Vektoren.
Die beste Accuracy wird in den Bereichen *Präteritum dritte Person Singular* und *Geschlecht* erreicht. Dies lässt darauf schließen, dass Worte aus diesen Bereichen oft in den Trainingsdaten vorkommen, sodass ein deutlicher Zusammenhang zwischen den Vektoren besteht.

4.2.4 Modell mit Wörterbuch aus Bewertungsdaten und Training mit Wikipediadaten (*Wiki_1*)

Beim dritten Modell wurde das Wörterbuch aus den Bewertungsdaten erstellt, wobei hier nur Worte betrachtet wurden, welche mindestens 5 mal in den Daten vorkommen. Dadurch ergibt sich eine deutlich höhere Abdeckung der Worte aus dem Fragenkatalog als beim reinen Wikipedia Modell und eine leicht höhere Abdeckung als beim Bewertungsdaten Modell.
Trainiert wurde das Modell mit den Bewertungen und den kompletten Wikipediadaten. Bei der Erstellung des Wörterbuches auch Worte mit einzubeziehen, welche mindestens 5 mal im Korpus vorkommen, hat den Vorteil, dass möglichst viele Worte aus dem Bereich Mode im Modell enthalten sind. Das zusätzliche Training mit den Wikipediadaten bringt fast die doppelte Genauigkeit gegenüber dem Modell, welches nur mit den Bewertungen trainiert wurde.
Aufgrund der großen Anzahl an Worten im Wikipediadatensatz und dem damit verbundenen Berechnungsaufwand wurden hier, anders als im reinen Bewertungsdaten Modell, keine Bigramme angewendet.
Die Abdeckung der Worte im Fragenkatalog beträgt 77%. Die gesamte Zahl der richtig beantworteten Fragen aus dem Fragenkatalog beträgt 28,5%. Dies sind weniger als

[3] http://de.wikipedia.org/wiki/Wikipedia, abgerufen am 13.01.2015

beim reinen Wikipedia Modell. Bezieht man jedoch die hohe Abdeckung mit ein, ist dies insgesamt trotzdem ein deutlich besseres Ergebnis, da absolut gesehen mehr Fragen aus dem Fragenkatalog richtig beantwortet wurden.

Tabelle 4.5: Accuracy Modell *Wiki_1*

Kategorie	Accuracy	
	[%]	absolut
Gegenteil	32,7	(89/272)
Perfekt erste Person Singular	14,3	(49/342)
Präteritum dritte Person Singular	41,1	(156/380)
Steigerung	39,8	(167/420)
Plural Nomen	17,4	(73/420)
Geschlecht	54,2	(147/271)
Superlativ	11,8	(13/110)
Präteritum erste Person Singular	12,1	(46/380)
total	28,5	(740/2595)

Um eine höhere Genauigkeit bei der Beantwortung der semantischen und syntaktischen Fragen zu erreichen, muss mit mehr Daten trainiert werden. Diese Daten sollten dabei aus dem selben Kontext wie die Bewertungsdaten stammen und somit möglichst viele der Worte aus dem Wörterbuch enthalten.
Wie der Vergleich des Modells aus den Bewertungsdaten und des Modells, welches zusätzlich mit den Wikipediadaten trainiert wurde, zeigt, hat die Menge der Trainingsdaten Einfluss auf die Genauigkeit. Dieser Einfluss hängt davon ab, ob und wie oft die Worte aus dem Fragenkatalog in den zusätzlich verwendeten Trainingsdaten vorkommen.
In [Mik+13a] wird für die englische Sprache eine Genauigkeit von 65,6% erreicht. Dieses Modell wurde mit dem Google News Datensatz trainiert, welcher mehr als 6 Milliarden Worte enthält.

4.2.5 Modell mit Wörterbuch aus Bewertungsdaten und Training mit Wikipediadaten 2 (*Wiki_2*)

Die Ergebnisse aus [Mik+13b] zeigen dass eine Verdopplung der Dimensionen der Wortvektoren in gewissem Rahmen einen ähnlichen Effekt auf die Genauigkeit bei der Beantwortung der Fragen aus dem Fragenkatalog hat wie die Verdopplung der Trainingsdaten. Daher wurde hier gegenüber dem Modell aus Kapitel 4.2.4 der Wert für den Parameter *size* verdoppelt. Die Verbesserung der Genauigkeit beträgt hierbei 2%. Dies lässt darauf schließen, dass eine weitere Steigerung der Dimensionen keine weitere große Optimierung bringen wird. An diesem Punkt ist eine signifikante Steigerung der Genauigkeit nur durch eine Erhöhung der Menge der Trainingsdaten realisierbar.

Tabelle 4.6: Accuracy Modell *Wiki_2*

Kategorie	Accuracy	
	[%]	absolut
Gegenteil	35,7	(97/272)
Perfekt erste Person Singular	14,9	(51/342)
Präteritum dritte Person Singular	45,3	(172/380)
Steigerung	41,9	(176/420)
Plural Nomen	18,3	(77/420)
Geschlecht	61,3	(166/271)
Superlativ	9,1	(10/110)
Präteritum erste Person Singular	11,3	(43/380)
total	30,5	(792/2595)

Auffällig ist, dass sich in allen Bereichen die Genauigkeit leicht steigert, außer in den Bereichen *Superlativ* und *Präteritum erste Person Singular*. Dies sind die Bereiche, welche schon im vorherigen Modell nur einen schwachen syntaktischen Zusammenhang zeigten.
Im Gegensatz dazu hat sich die Genauigkeit in den Bereichen *Geschlecht* und *Präteritum dritte Person Singular* im Vergleich zum vorherigen Modell mit 6,9% und 4,2% deutlich gesteigert. Dies deutet darauf hin, dass die semantischen und syntaktischen Zusammenhänge, welche bei weniger Dimensionen schon deutlich waren, durch die Steigerung der Dimensionen weiter gestärkt werden. Zusammenhänge, die jedoch bei weniger Dimensionen schon eher schwach ausgeprägt sind, werden durch die Steigerung der Dimensionen noch geringer.

4.2.6 Modell mit Wörterbuch aus Bewertungsdaten und Training mit Wikipediadaten 3 (*Wiki_3*)

Um zu zeigen wie sich der semantische und syntaktische Zusammenhang bei Worten verhält, welche nur selten in den Trainingsdaten vorkommen wurde ein Modell erstellt, bei welchem die minimale Worthäufigkeit 10 beträgt.
Wie beim Modell in Kapitel 4.2.5 wurde hier der Parameter *size* auf einen Wert von 600 eingestellt, keine Berechnung der Bigramme vorgenommen und das Wörterbuch wurde mit Hilfe der Bewertungsdaten erstellt. Trainiert wurde auch hier mit den Bewertungsdaten und dem kompletten Satz der Wikipediadaten.

Tabelle 4.7: Accuracy Modell *Wiki_3*

Kategorie	Accuracy	
	[%]	absolut
Gegenteil	34,2	(93/272)
Perfekt erste Person Singular	12,1	(33/272)
Präteritum dritte Person Singular	47,7	(163/342)
Steigerung	45,0	(189/420)
Plural Nomen	17,9	(75/420)
Geschlecht	56,9	(119/209)
Superlativ	13,3	(4/30)
Präteritum erste Person Singular	11,1	(38/342)
total	30,9	(714/2307)

In Tabelle 4.7 zeigt sich, dass durch die Anhebung der minimalen Worthäufigkeit die Abdeckung von 77% auf 69% abfällt. Das bedeutet, dass einige der Worte aus dem Fragenkatalog in den Trainingsdaten nur selten vorkommen. Wie erwartet ist dies vor allem bei den Bereichen *Superlativ* und *Perfekt erste Person Singular* der Fall, welche in den Modellen *Wiki_1* und *Wiki_2* auch schon eine eher niedrige Genauigkeit erzielten.
Entgegen der Erwartungen jedoch steigt die Genauigkeit insgesamt nur minimal an. Dies lässt darauf schließen, dass auch bei Worten, welche nur selten in den Trainingsdaten vorkommen, ein messbarer syntaktischer Zusammenhang erzeugt werden kann.
Auch beim semantischen Zusammenhang der Worte aus dem Bereich Geschlecht zeigt sich ein überraschendes Ergebnis. Die gesamte Anzahl der durch das Wörterbuch abgedeckten Fragen sinkt um 62 Fragen, wobei gleichzeitig auch die Anzahl der durch das Modell korrekt beantworteten Fragen um 47 Fragen sinkt. Dies deutet darauf hin, dass nicht nur der syntaktische, sondern auch der semantische Zusammenhang zwischen Worten, die nur selten in den Trainingsdaten vorkommen, errechnet werden kann.

4.2.7 Zusammenfassung

In der Summe lässt sich sagen, dass die hier gezeigten Word2Vec Modelle gute Resultate liefern. Die Ergebnisse lassen sich nur bedingt mit den Ergebnissen aus [Mik+13b] vergleichen, da eine andere Sprache und auf Grund der unterschiedlichen Sprache auch ein anderer Trainingsdatensatz sowie ein anderer Fragenkatalog verwendet wurden. Um die verschiedenartigen Zusammenhänge, welche mit Word2Vec erfasst werden können, auch außerhalb der Genauigkeit bei der Beantwortung der Fragen aus dem Fragenkatalog darzustellen, wurden die Vektoren im nachfolgenden Kapitel in Cluster eingeteilt. Diese Cluster geben interessante Einblicke in die Möglichkeiten, welche sich mit Word2Vec bieten.
Das Modell *Wiki_2* erzielt im Hinblick auf Abdeckung und gesamte Genauigkeit bei der Beantwortung der Fragen aus dem Fragenkatalog von allen Modellen die besten Ergebnisse und wird daher nachfolgend für das Clustering, sowie auch für die Merkmalsextraktion genutzt.

4.3 K-means Cluster mit Word2Vec Modell *Wiki_2*

Um die starken inhaltlichen Zusammenhänge zwischen den ähnlichen durch Word2Vec gelernten Vektoren zu verdeutlichen wurden die Vektoren mit Hilfe des K-Means Algorithmus in Cluster unterteilt. Mit K-Means konnten bei [TW14] in Verbindung mit Word2Vec Modellen gute Ergebnisse erzielt werden. Das K-Means Clustering wird auch von Mikolov et al. vorgeschlagen[4].
Einige Besonderheiten der erhaltenen Cluster werden im Anschluss vorgestellt. Die Cluster dienen nur der Veranschaulichung der bisherigen Ergebnisse und werden später nicht weiterverwendet.

4.3.1 K-Means Algorithmus

Der K-Means Algorithmus ist eines der bekanntesten Clusterverfahren. Hierbei werden die Vektoren zu einer Menge von k Clustern zugeordnet. Die Zahl k der zu berechnenden Cluster wird von Hand festgelegt.
Die Zuordnung erfolgt hart, was bedeutet, dass ein Vektor genau einem Cluster zugeordnet wird [Ert09]. K-Means wird hier zur Veranschaulichung der Zusammenhänge zwischen den Vektoren verwendet, da er sehr schnell ist und trotzdem gute Ergebnisse liefert, mit welchen die Zusammenhänge in den Vektoren veranschaulicht werden können.
Die richtige Wahl für die Größe des Parameters k zu wählen ist nicht einfach. Die obere Grenze für k definiert sich über die Anzahl der Wortvektoren im verwendeten Word2Vec Modell. Im vorliegenden Fall wurde die Clusteranzahl so gewählt, dass ein Cluster im Durchschnitt 5 Worte enthält. Da die erhaltenen Cluster die bisherigen Annahmen soweit unterstützen, die Berechnung der Cluster, sowie die Zuweisung der Worte zu den Clustern zur Ansicht sehr lange dauert[5] und die Cluster nicht für weitere Verarbeitungsschritte genutzt werden, wird kein Test mit einer anderen Anzahl von Clustern durchgeführt. Wenn die Cluster zum Beispiel zur Gruppierung der Merkmale genutzt werden sollen, so müssen Tests mit mehreren verschiedenen Werten für k durchgeführt werden sowie eine geeignete Metrik zur Überprüfung der Qualität der Cluster angewendet werden.

4.3.2 Clustering

Die gelernten Vektoren des Modells *Wiki_2* wurden mit Hilfe des K-Means Algorithmus in 4753 Cluster unterteilt.
Etwas mehr als 4000 der Cluster enthalten nur ein Wort. Die übrigen Cluster enthalten je eine Gruppe von Worten mit unterschiedlichsten Zusammenhängen. Ein Teil dieser Cluster wird im Folgenden näher beschrieben und der Zusammenhang verdeutlicht.

Synonyme
Ein Teil der Cluster bildet unterschiedliche Schreibweisen eines Wortes ab.
Cluster 187
[u'preis-leistungs-verh\xe4ltnis', u'preis/leistungsverh\xe4ltnis', u'preisleistungsverh\xe4ltnis', u'preis-/leistungsverh\xe4ltnis', u'preis-leistungsverh\xe4ltnis']

[4] https://code.google.com/p/word2vec/, abgerufen am 28.01.2015
[5] mehr als 12h auf einem MacBook Pro mit 8GB Ram und 2 Kernen mit je 2,3GHz

Cluster 1275
[u'preis/leistung', u'preis-leistung', u'verh\xe4ltniss', u'preis-leistungs', u'preisleistung', u'preisverh\xe4ltnis', u'preisleistungs', u'preisleistungsverh\xe4ltniss', u'preis-leistung-verh\xe4ltnis', u'preis-leistungsverh\xe4ltniss']
Cluster 2147
[u'dekolette', u'dekollet\xe8', u'dekolltee', u'dekollte', u'dekollt\xe9', u'dekolt\xe9', u'dekolte', u'dekolt\xe9e', u'dekolte\xe9', u'dekolletee']

Verben
Die folgenden Cluster enthalten unterschiedlichen Konjugationen eines Verbes.
Cluster 2722
[u'geeignet', u'eignen', u'eignet']
Cluster 3025
[u'erweisen', u'erwiesen', u'erweist', u'erwies']

Gegenteile
Cluster 3143
[u'helles', u'dunkles']
Cluster 3513
[u'weichem', u'hartem']
Cluster 3704
[u'nachteil', u'vorteil']

Semantische Zusammenhänge
Einige der Cluster bilden komplexere semantische Zusammenhänge ab, wie zum Beispiel unterschiedliche Materialien, welche zur Herstellung von Kleidung verwendet werden, Monate und Worte aus dem Bereich Sport und Fitness.
Cluster 626
[u'kunstfaser', u'seide', u'nylon', u'baumwolle']
Cluster 1572
[u'baumwollgewebe', u'elasthan', u'schurwolle', u'leder', u'elastan', u'haltbarer', u'baumwollanteil', u'stoffe', u'obermaterial', u'kunstfasern', u'lycra', u'naturfaser', u'stoffen', u'polyacryl', u'synthetischer', u'synthetischen', u'elastischer', u'baumwoll', u'gewebt', u'gewebe', u'viskose', u'mischgewebe', u'frottee', u'mikrofaser', u'baumwollstoff', u'synthetik']
Cluster 92
[u'november', u'august', u'dezember', u'april', u'm\xe4rz', u'september', u'februar', u'mai', u'oktober', u'januar', u'juli', u'juni']
Cluster 454
[u'ballsport', u'fitnesstraining', u'wassergymnastik', u'krafttraining', u'skifahren', u'ausdauertraining', u'ausdauersport', u'freizeitsport', u'pilates', u'jogging', u'laufsport', u'schulsport']

Verben aus dem Bereich „Bewegung“
Eines der größeren Cluster bildet den semantischen Zusammenhang zwischen verschiedenen Verben aus dem allgemeinen Bereich „Bewegung“ ab. Bedenkt man den Kontext, aus welchem die Daten stammen, so ist dies ein sehr erstaunliches Ergebnis. Mit nur vergleichsweise wenigen Daten können hier komplexe Zusammenhänge erkannt werden.
Bewegung Cluster 2289

[u'rutschen', u'gleitet', u'wackeln', u'gerutscht', u'blitzschnell', u'abgelenkt', u'hochgezogen', u'schleift', u'hintern', u'h\xfcpfen', u'presst', u'lockert', u'bewegt', u'dreht', u'gelegte', u'versinkt', u'rei\xdft', u'freigibt', u'verfangen', u'schleudern', u'ber\xfchrt', u'eingeklemmt', u'flattert', u'rollt', u'springt', u'herunter', u'rasend', u'reisst', u'schiebt', u'zuschl\xe4gt', u'verf\xe4ngt', u'wickelt', u'umgedreht', u'h\xfcpft', u'geschoben', u'nachgibt', u'hin\xfcber', u'rutscht', u'rafft', u'sitzt', u'wandert', u'b\xfcckt', u'ungewollt', u'pickt', u'faltet', u'dr\xfcckt', u'kippt', u'beugt', u'gedr\xfcckt', u'aufsteht', u'bohrt', u'kriechen', u'zusammenzieht', u'hakt', u'versinken', u'schwebt', u'klemmt', u'st\xfclpt', u'anhebt', u'aufsetzt', u'zerrei\xdft', u'l\xe4uft']

Abkürzungen
Auch das Erkennen des semantischen Zusammenhangs von Abkürzungen ist ein interessantes Ergebnis, welches für vielfältige Aufgaben, zum Beispiel zur Satzsegmentierung (siehe Kapitel 2.6.3) genutzt werden kann um zu verhindern, dass Sätze nach Abkürzungen fälschlicherweise getrennt werden.
Cluster 3638
[u'vgl', u'sog', u'bzw', u'ggfs', u'vllt', u'ggf', u'usw', u'jmd', u'evt', u'evtl', u'vll', u'vlt', u'bspw', u'etc', u'\xe4']

Farben
Der semantische Zusammenhang zwischen Farben ist besonders stark ausgeprägt. Es finden sich einige Cluster, die ausschließlich Farben enthalten.
Cluster 1415
[u'taupe', u'limone', u'rauchblau', u'petrol', u'wollwei\xdf', u'rinsed', u'apricot', u'mauve', u'royalblau', u'kaki', u'aubergine']
Cluster 3537
[u'dunkelrot', u'f\xe4rbt', u'fahl', u'rostrot', u'meliert', u'smaragdgr\xfcn', u'hellrot', u'lila', u'goldfarben', u'schimmert', u'grasgr\xfcn', u'ocker', u'lachsfarben', u'zitronengelb', u'graublau', u'weinrot', u'braun', u'schmutzig']

Es zeigt sich, dass durch das Clustering der Vektoren aus dem Word2Vec Modell *Wiki_2* viele semantische und syntaktische Zusammenhänge zwischen den Worten dargestellt werden können. Einige dieser Zusammenhänge sind überraschend und können auch in anderen Kontexten als dieser Arbeit praktisch genutzt werden.

4.4 Merkmale annotieren

Um mit Word2Vec Merkmale aus den Bewertungen zu finden ist es notwendig zuerst einige der Bewertungen von Hand zu annotieren. Da Word2Vec nur semantische und syntaktische Zusammenhänge findet, müssen zuerst Merkmale gesammelt werden um mit Hilfe des Word2Vec Modells ähnliche Merkmale zu finden.
Bei der Annotation der Bewertungen werden Worte, welche als Merkmale eines Produktes genannt wurden, markiert und in einem Set gespeichert.
Es wurde ein Set von 10000 zufälligen Sätzen aus den Bewertungen aus dem Bereich der Damen- und Herrenmode ausgewählt und manuell annotiert. Dabei wurden nicht die durch Rechtschreibkorrektur verbesserten Daten, sondern die Rohdaten verwendet, da auch die Word2Vec Modelle mit den nicht korrigierten Daten trainiert wurden.

Es wurden nicht nur die Merkmale, sondern auch die dazugehörigen Meinungsworte markiert. Bei den Meinungsworten wurde zusätzlich mit der Information annotiert, ob diese eine positive oder negative Stimmung gegenüber dem bewerteten Merkmal ausdrücken.
Wie in [Pon+14] vorgeschlagen wurde zur Annotation das Programm *Brat* (siehe Kapitel 2.3) genutzt, mit welchem sich einfach und schnell größere Mengen an Textdaten annotieren lassen.

4.4.1 Merkmalsliste erweitern

Aus den manuell annotierten Bewertungen lassen sich 221 Merkmale extrahieren (siehe Anhang B.2). Diese enthalten meist viele Versionen eines Merkmals, da diese in den Bewertungen oft falsch geschrieben werden. Auch mit Hilfe der Rechtschreibkorrektur können nicht alle Versionen auf die korrekte Schreibweise zurückgeführt werden. In Kapitel 4.5 wird dies näher erläutert.

Um ein erweitertes Set von Merkmalen zu erhalten wurden zu den 221 manuell annotierten Merkmalen mit Hilfe von Word2Vec die dazu ähnlichsten Wortvektoren berechnet.
Das Word2Vec Modell, welches für diese Berechnungen genutzt wurde, ist in Kapitel 4.2.5 beschrieben.
Die Ergebnisse zur Erweiterung der Merkmalsliste sind in Tabelle 4.8 dargestellt. Es wurden verschieden Ähnlichkeitswerte getestet. Dabei wurden zunächst zu jedem Wort aus der Liste der manuell annotierten Merkmale die Worte gesucht, welche eine Ähnlichkeit (siehe Kapitel 4.1.6) von minimal dem jeweils eingestellten Wert aufweisen.
Dabei fiel auf, dass einige der gefundenen Worte Größenangaben wie z.B. 1,75m oder 180cm sind. Diese können leicht durch den regulären Ausdruck in Formel 4.4 gefiltert werden.

$$\wedge\,[0-9,]+(cm|m)\$ \tag{4.4}$$

Zusätzlich werden die entstandenen Listen noch nach Kategorien, Marken und Meinungsworten (siehe Kapitel 2.1.2) gefiltert.
Um möglichst viele Worte zu filtern, welche keine Merkmale darstellen, wurde zusätzlich jedes Wort mit der Rechtschreibkorrektur geprüft. Taucht das Wort nach der Korrektur in den Marken, Kategorien oder Meinungsworten auf, so wird es ebenfalls aus der Liste gelöscht.

Tabelle 4.8 zeigt die Ergebnisse der erweiterten Merkmalslisten mit verschiedenen Werten für die minimale Kosinus Distanz zwischen einem der Vektoren aus der Liste der manuell annotierten Worte und einem Wort aus dem Word2Vec Modell. Es werden die Ergebnisse jeweils vor und nach der Filterung, sowie nach der manuellen Durchsicht gezeigt.
Bei Ähnlichkeit 0.75 gegenüber einem der Merkmale aus der manuell annotierten Liste sind nur 2 dieser Worte keine Merkmale. Jedoch steigert sich die Menge der Merkmale gegenüber der Ursprungsliste nur um 30%.
Das beste Verhältnis zwischen falsch positiven Merkmalen und der Steigerung der gesamten Anzahl der Merkmale gegenüber der Ursprungsliste ergibt sich bei einer Ähnlichkeit von 0.65. Hierbei werden nur 16.6% der Worte falsch als Merkmale deklariert,

Tabelle 4.8: erweiterte Merkmalslisten

Ähnlichkeit	total	gefiltert	nach Durchsicht	% der gefilterten
0.75	354	288	286	99,3%
0.70	506	377	345	91,5%
0.65	773	602	502	83,4%
0.60	1286	1007	734	72,9%
0.55	1775	1385	884	63,8%

die Ursprungsliste erweitert sich jedoch um 172%.
Bei manueller Durchsicht fallen typischerweise Adjektive wie *verwaschenen* und *topmodische* oder Nomen, welche die Produkte selbst bezeichnen wie *Frühlingsbluse* und *Business-Hose* auf, welche fälschlicherweise als Merkmale markiert wurden.
Bei der Erweiterung der Merkmalsliste wird deutlich, warum das Word2Vec Modell mit allen Bewertungsdaten trainiert wurde. Um später evaluieren zu können, ob durch die Erweiterung zur Merkmalsliste auch Merkmale aus dem Bereich Multimedia hinzugekommen (siehe Kapitel 7) sind, ist es wichtig, dass die Worte aus den Bewertungen aus dem Bereich Multimedia auch im Wörterbuch des Word2Vec Modells vorkommen. So können potentiell domänenfremde Merkmale in die erweiterte Merkmalsliste gelangen.

4.5 Evaluation der Merkmalsextraktion mit Word2Vec

Wie schon im Kapitel 4.3.2 gezeigt können mit Word2Vec sehr gut Synonyme gefunden werden. Ebenso wie verschiedene Schreibweisen eines Wortes, welche auch Synonyme darstellen.
Wird die minimale Kosinus Distanz niedriger eingestellt, so werden mit Word2Vec nicht nur die Synonyme gefunden, sondern auch andere semantisch zusammengehörige Worte, wie zum Beispiel beim Cluster mit verschiedenen Materialien.
Durch das leichte Auffinden von Synonymen eines Wortes ist es nicht notwendig die Daten vor der Berechnung der Word2Vec Modelle mittels Rechtschreibkorrektur zu korrigieren. Dies ist von Vorteil, da falsche Schreibweisen von Worten nicht durch die Rechtschreibkorrektur, wie sie hier implementiert ist, korrigiert werden können. Nimmt man 20 zufällige Versionen des Wortes *Preisleistungsverhältnis* (siehe Anhang Tabelle B.3) aus der erweiterten Merkmalsliste, so zeigt sich, dass sich nur 12 der 20 Versionen mittels Rechtschreibkorrektur auf das korrekte Wort zurückführen lassen. Bei 20 Versionen des Wortes *Qualität* sind es 14, welche korrekt korrigiert werden (siehe Anhang Tabelle B.4). Dabei ist anzumerken, dass die Worte jeweils in Kleinschreibweise verwendet wurden.
Um mit Word2Vec Merkmale zu extrahieren ist es jedoch notwendig zuerst eine Liste von tatsächlichen Merkmalen in der Domäne zu erstellen. Je umfangreicher diese Liste ist, desto mehr Merkmale können durch die Merkmalserweiterung errechnet werden. Dabei ist es wichtig die minimale Kosinus-Distanz so zu setzen, dass die Merkmalsliste deutlich erweitert wird, ohne dass zu viele Worte fälschlicherweise als Merkmale in die Liste mit aufgenommen werden.
In dieser Arbeit wurden die Word2Vec Modelle mit einer relativ geringen Datenmenge

trainiert. Es ist zu erwarten, dass sich die Ergebnisse nach dem Training mit mehr Daten verbessern, insbesondere wenn diese Daten aus der gleichen Domäne (Mode) stammen und im besten Fall auch Bewertungen sind.

Kapitel 5

Vergleich der Methoden zur Merkmalsextraktion

Im folgenden Abschnitt werden die Ergebnisse der beiden Methoden häufigste Nomen und Word2Vec zur Extraktion von Merkmalen aus Bewertungen verglichen. Anschließend werden die Vor- und Nachteile der beiden Methoden herausgearbeitet.

5.1 Vergleich der Ergebnisse

Vergleicht man die Ergebnisse der Merkmalsextraktion durch häufigste Nomen mit der handgetaggten Liste der Merkmale, sowie der erweiterten Merkmalsliste (mit minimaler Kosinus Distanz 0.65) so erkennt man deutliche Unterschiede.
Von den 221 Worten in der handgetaggten Liste sind nur 51 in der Liste der häufigsten Nomen enthalten. Von der erweiterten Merkmalsliste sind es nur wenig mehr; genauer 56.
Trotz der Rechtschreibkorrektur sind bei den häufigsten Nomen teils noch unterschiedliche Versionen eines Wortes vorhanden, da die Rechtschreibkorrektur nicht alle Fehlschreibweisen erkennt (wie zum Beispiel im Anhang in den Tabellen B.3 und B.4 veranschaulicht).

5.2 Vor- und Nachteile der Methoden

Der größte Nachteil bei der Merkmalsextraktion mittels der häufigsten Nomen besteht darin, dass diese stark von der Qualität der Daten abhängt, beziehungsweise auch davon, ob die Daten durch die Vorverarbeitung die nötige Qualität erreichen.
Ein weiterer Punkt ist die Tatsache, dass die häufigsten Worte erst gefiltert werden müssen, bevor sie überhaupt als Merkmalsliste in Frage kommen. Auch wenn die Marken und Kategorienamen herausgefiltert sind, bleiben noch Nomen, welche keine Merkmale darstellen. Dadurch ergibt sich, dass die Liste manuell nachgearbeitet werden muss, bevor sie tatsächlich auf neuen Daten angewendet werden kann.
Seltene Merkmale können mit dieser Methode überhaupt nicht gefunden werden.
Auch der zeitliche Aufwand durch das Part-of-Speech Tagging ist ein Nachteil dieser Methode.

Die Merkmalsextraktion mittels Word2Vec hängt sehr stark von der Ausgangsliste ab. Ohne eine Ausgangsliste von handgetaggten Merkmalen können keine neuen Merkmale gefunden werden.
Es werden abhängig von der Ausgangsliste teils auch Merkmale gefunden, welche bei einem Produkt zwar Merkmale beziehungsweise Bestandteile bezeichnen, bei einem anderen Produkt aber den Artikel an sich bezeichnen und somit kein Merkmal darstellen. Ein Bespiel hierfür ist das Wort *Weste*. Dies kann sowohl ein einzelnes Produkt sein, als auch Bestandteil eines dreiteiligen Anzugs.
Ein Vorteil besteht darin, dass es beim manuellen Annotieren leichter ist zu erkennen, ob es sich tatsächlich um ein Merkmal handelt, als bei der Liste der häufigsten Nomen, da man dort den Kontext der gesamten Bewertung zur Verfügung hat.
Ein weiterer Vorteil von Word2Vec ist, dass hierbei verschiedene Schreibweisen eines Wortes leicht gefunden werden können. Wie am Beispiel des Wortes *Preis-Leistungs-Verhältnis* gezeigt, enthält die Liste der häufigsten Nomen trotz Rechtschreibkorrektur noch mehrere Versionen des Wortes. Mit Word2Vec kann der Zusammenhang zwischen den verschiedenen Versionen eines Wortes abgebildet werden, wie beim Clustering in Kapitel 4.3.2 veranschaulicht ist.
Ebenso von Vorteil ist, dass im Vergleich zum Part-of-Speech Tagging die Berechnung der Word2Vec Modelle sehr schnell ist. Selbst mit einer großen Menge von Trainingsdaten kann innerhalb weniger Stunden bis Tage ein neues Word2Vec Modell berechnet werden.

Ein Nachteil beider Methoden ist es, dass keine Merkmale gefunden werden, welche aus mehr als einem Wort bestehen. Bei der Merkmalsextraktion mittels Word2Vec ist dies grundsätzlich durch die Berechnung von Bigrammen oder auch Trigrammen (Trigramme sind Gruppen von drei Worten, welche häufig zusammen auftreten) möglich. Jedoch erfordert dies auch eine deutlich höhere Berechnungszeit und eine große Menge an Trainingsdaten, um die Bigramme und Trigramme zuverlässig berechnen zu können.

Ein weiterer Nachteil beider Methoden ergibt sich wenn neue Bewertungen hinzugefügt werden. Insbesondere wenn diese Bewertungen neue Merkmale enthalten. Um mittels der Methode der häufigsten Nomen diese neuen Merkmale extrahieren zu können müssen diese schon sehr oft genannt sein. Nach dem Hinzufügen neuer Bewertungen muss gegebenenfalls auch die Grenze ab welcher Häufigkeit ein Wort ein Merkmal ist verändert werden.
Bei Word2Vec muss bei neuen Merkmalen das verwendete Modell neu berechnet werden, da ein Modell zwar mit mehr Worten trainiert werden kann, jedoch das Wörterbuch nicht nachträglich erweitert werden kann.

Kapitel 6

Merkmalsbasiertes Opinion Mining

In diesem Kapitel wird nun gezeigt, wie mit Hilfe der gefundenen Merkmale eine merkmalsbasierte Stimmungsanalyse realisiert werden kann. Da durch die Extraktion potentieller Merkmale mittels der manuell annotierten Liste und der Erweiterung dieser Liste durch Word2Vec deutlich bessere Ergebnisse erzielt werden, wird die erweiterte Merkmalsliste aus Kapitel 4.4.1 verwendet (siehe auch im Anhang B.5). Diese wurde mit einer Ähnlichkeit der Worte von mindestens 0.65 gegenüber einem der Worte aus der manuell annotierten Merkmalsliste erstellt. Eine Menge von 16.6% falsch positiv annotierter Merkmale gegenüber einer Steigerung der Menge an Merkmalen von 172% der Ursprungsliste wird dabei als vertretbar erachtet.
Im Folgenden werden die Daten, welche für den Test des merkmalsbasierten Opinion Minings genutzt wurden, sowie der Ablauf näher beschrieben.

6.1 Daten

Für den Test des merkmalsbasierten Opinion Minings wurden 250 Bewertungen herangezogen, welche nicht zum Training des Word2Vec Modells genutzt wurden. Davon sind 100 Bewertungen aus zufälligen Kategorien des Bereichs Damenmode, 100 aus dem Bereich Herrenmode, sowie 50 aus dem Bereich Multimedia.
Anhand der Ergebnisse mit den Daten aus dem Bereich Multimedia wird bewertet wie sich die Merkmalsextraktion bei Daten verhält, welche nicht aus dem Bereich Mode stammen.
Da das Word2Vec Modell auch mit Bewertungen aus dem Bereich Multimedia trainiert wurde, sollten die Worte, welche im Bereich Multimedia Merkmale darstellen, im Wörterbuch vorhanden sein. Da zur manuellen Annotation nur Daten aus den Kategorien Damen- und Herrenmode betrachtet wurden, ist es interessant zu betrachten, ob in der Liste der erweiterten Merkmale auch Merkmale aus dem Bereich Multimedia vorhanden sind.

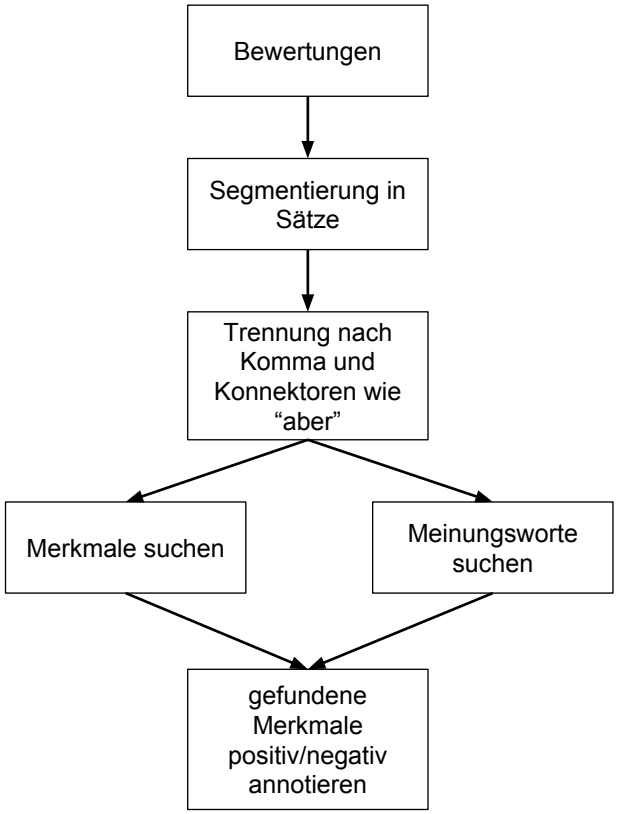

Abbildung 6.1: Ablauf des Opinion Minings

6.2 Ablauf

Angesichts der Qualität der Ursprungsdaten wurde für das merkmalsbasierte Opinion Mining ein sehr simpler Ansatz gewählt. Der Ablauf ist auch in Abbildung 6.1 dargestellt.
Zunächst wurden die Bewertungen in Sätze unterteilt. Daraufhin wurden sie an Kommatas getrennt. Taucht in einem so entstandenen Satzfragment ein Konnektor[1] auf, der darauf hindeutet, dass sich die Meinung innerhalb dieses Satzfragmentes umkehrt, so wird an dieser Stelle erneut getrennt. Diese Konnektoren sind konkret die adversativen (zum Beispiel *aber*, *allerdings*, *während*) und die konzessiven Konnektoren (zum Beispiel *doch*, *jedoch*, *obgleich*)[2]. In der Version des Programmes, welches für die Bewertung in Kapitel 7 genutzt wurde, wurde nur nach dem Konnektor *aber* getrennt.
In den entstandenen Satzfragmenten wurden daraufhin die Merkmale aus der Merkmalsliste markiert, sowie nach Meinungsworten gesucht.

[1] http://hypermedia.ids-mannheim.de/call/public/gramwb.ansicht?v_app=g&v_kat=Konnektor, abgerufen am 24.01.2015

[2] http://hypermedia.ids-mannheim.de/call/public/sysgram.ansicht?v_typ=d&v_id=366, abgerufen am 24.01.2015

Zur Erkennung der Stimmungen zu den jeweiligen Merkmalen wurde SentiWS [RQH10] genutzt. Dies ist eine Sammlung von Worten mit positiver und negativer Konnotation. Zusätzlich dazu wurden die Meinungsworte aus den manuell annotierten Bewertungen verwendet, da diese unter anderem auch umgangssprachliche Worte (wie zum Beispiel *altbacken* oder *trendigen*) enthalten, oder spezielle Meinungsworte, welche nur im Zusammenhang mit Bewertungen im Bereich Mode verwendet werden (zum Beispiel *kastig* oder *anschmiegsam*).
Kommt in einem Satz ein Meinungswort zusammen mit einem Merkmal vor, so wird das Merkmal mit +1, wenn es sich um ein positives Meinungswort handelt, oder -1, wenn es sich um ein negatives Meinungswort handelt, annotiert. Taucht zusätzlich zum Meinungswort eine Negation auf (wie zum Beispiel *nicht* oder *kein*) wird der Wert der Annotation umgekehrt.
Die Stimmungen werden nicht in Stufen dargestellt, sondern die positive oder negative Tendenz im Bezug auf ein Merkmal erfasst.
Es wurden keine Meinungen betrachtet, welche aus mehr als einem Wort bestehen. Es wurden lediglich einzelne Meinungsworte und Meinungsworte in Verbindung mit einem Negationswort betrachtet.

Kapitel 7

Bewertung der Ergebnisse aus dem merkmalsbasierten Opinion Mining

In diesem Kapitel wurden die Ergebnisse aus dem merkmalsbasierten Opinion Mining durch zwei unabhängige Gruppen von Personen bewertet. Es wurde jeweils eine Bewertung der Qualität der Merkmalsextraktion, sowie eine Bewertung der Sentimentanalyse durchgeführt.
Die dazu genutzte Vorgehensweise wird erläutert und die Resultate beschrieben.

7.1 Vorgehensweise

Die Merkmalsextraktion, sowie die Stimmungsanalyse dieser Merkmale, wurden von zwei unabhängigen Gruppen bewertet. Jede dieser Gruppen bestand aus vier Personen. Diese bekamen jeweils die Ergebnisse der merkmalsbasierten Stimmungsanalyse von 100 Bewertungen der Kategorie Damenmode, 100 Bewertungen der Kategorie Herrenmode, sowie 50 Bewertungen der Kategorie Multimedia vorgelegt. Die einzelnen Kundenbewertungen wurden den Gruppen wie in Abbildung 7.1 dargestellt gezeigt.
Die in den Kundenbewertungen gefundenen Merkmale wurden durch die automatische Merkmalsextraktion mittels der erweiterten Merkmalsliste, sowie der Stimmungsanalyse jeweils in positiv, negativ und neutral bewertete Merkmale eingeteilt und in die jeweilige Spalte eingetragen. Die zwei Gruppen bearbeiteten daraufhin unabhängig von-

Bewertung	positiv	negativ	neutral
Super Mantel, schöner Schnitt und Farbe. Allerdings schon etwas übertrieben der Preis.	Schnitt, Farbe	Preis	-

Abbildung 7.1: Beispiel einer Kundenbewertung mit Merkmalsextraktion und Stimmungsanalyse

einander jeweils für die gleichen 250 Kundenbewertungen folgende Punkte:

- Wurden zu viele Merkmale gefunden und wenn ja wie viele?
- Wurden zu wenige Merkmale gefunden und wenn ja wie viele?

- Zu wie vielen der korrekt gefundenen Merkmalen wurde die Stimmung richtig angegeben?
- Zu wie vielen der korrekt gefundenen Merkmale wurde die Stimmung falsch angegeben?

Die beiden Gruppen erhielten dabei die Anweisung nur explizite Merkmale zu betrachten.

Um die Güte der Merkmalsextraktion zu bewerten, wurden die Standardmaße im Gebiet des Information Retrieval *Precision* (*Genauigkeit*), *Recall* (*Vollständigkeit*) und das *F-Score* (*F-Maß*), welches sich aus den Werten für Precision und Recall errechnet, verwendet.
Precision beschreibt dabei wie viele der erkannten Merkmale tatsächlich Merkmale sind. Berechnet wird dies wie in Formel 7.1 gezeigt.

$$Precision = \frac{\Sigma\ richtig\ erkannte\ Merkmale}{\Sigma\ richtig\ erkannte\ Merkmale + \Sigma\ falsch\ erkannte\ Merkmale} \tag{7.1}$$

Der Recall beschreibt die Sicherheit mit der ein in der Bewertung vorhandenes Merkmal tatsächlich als solches erkannt wird. Die Berechnung des Recalls ist in Formel 7.2 dargestellt.

$$Recall = \frac{\Sigma\ richtig\ erkannte\ Merkmale}{\Sigma\ richtig\ erkannte\ Merkmale + \Sigma\ nicht\ erkannte\ Merkmale} \tag{7.2}$$

Aus Precision und Recall errechnet sich dann der F-Score (Formel 7.3).

$$F\text{-}Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{7.3}$$

Für die Bewertung der Stimmungsanalyse wurde errechnet wie viele der richtig erkannten Merkmale mit der korrekten Stimmung annotiert wurden.

7.2 Evaluation der Bewertungen

Im den folgenden zwei Abschnitten werden die Ergebnisse der Merkmalsextraktion und der Stimmungsanalyse untersucht und mit Ergebnissen aus anderen Quellen in Relation gestellt.

7.2.1 Merkmalsextraktion

Die Ergebnisse der Merkmalsextraktion sind in Tabelle 7.1 dargestellt. Darin ist jeweils für die Kategorien Damenmode, Herrenmode und Multimedia die Ergebnisse je Gruppe und im Durchschnitt Werte für die Precision, den Recall und den F-Score dargestellt.
Aus der Tabelle wird ersichtlich, dass bei allen Kategorien die Precision sehr gut ist. Der geringste Wert im Durchschnitt beider Gruppen liegt immer noch bei einer Precision

Tabelle 7.1: Ergebnisse der Merkmalsextraktion

Kategorie	Gruppe	Precision	Recall	F-Score
Damenmode	Gruppe 1	0,98	0,68	0,80
	Gruppe 2	0,99	0,73	0,84
	Schnitt	0,99	0,71	0,82
Herrenmode	Gruppe 1	0,96	0,89	0,92
	Gruppe 2	0,95	0,79	0,86
	Schnitt	0,96	0,84	0,89
Multimedia	Gruppe 1	1.00	0,25	0,40
	Gruppe 2	0,96	0,35	0,51
	Schnitt	0,98	0,30	0,46

von 0,96. Dies bedeutet, dass nahezu alle gefundenen Merkmale tatsächlich Merkmale sind.
Angesichts der Tatsache, dass in der erweiterten Merkmalsliste 16.6% der Worte keine Merkmale darstellen ist dieses Ergebnis sehr gut. Auch die im Bereich Multimedia gefundenen Merkmale haben eine sehr hohe Precision von 0,98.

Die Werte für den Recall sind deutlich niedriger, aber in den Kategorien Damen- und Herrenmode immer noch gut. Der Recall zeigt in diesem Fall wie viele der in den Bewertungen vorhandenen Merkmale durch die Merkmalsextraktion gefunden wurden.
Es überrascht nicht, dass die Werte für den Recall im Bereich Multimedia deutlich niedriger sind als für die Bereiche Damen- und Herrenmode. Da für die Ausgangsliste der erweiterten Merkmalsliste nur Bewertungen aus dem Bereich Mode annotiert wurden enthielt die Ausgangsliste nur Merkmale aus dem Bereich Mode. Es ist also anzunehmen, dass durch die Erweiterung der Merkmalsliste mittels Word2Vec keine oder kaum Merkmale aus der fremden Domäne zur Liste hinzugefügt werden konnten.
Sieht man sich die Ergebnisse der automatischen Merkmalsextraktion, welche den Gruppen zur Bewertung vorgelegt wurde an, bestätigt sich diese Annahme. Merkmale, welche im Bereich Multimedia korrekt automatisch gefunden wurden sind solche, welche für beide Domänen Mode und Multimedia valide Merkmale darstellen. Dies sind zum Beispiel *Qualität* und *Preis*.

Aus Precision und Recall ergibt sich der F-Score. Dieser ist bei den Kategorien Damen- und Herrenmode, durch die guten Werte in Precision und Recall ebenfalls hoch. Im Bereich Multimedia liegt der Wert durch den niedrigen Recall nur bei im Schnitt 0,46.

7.2.2 Stimmungsanalyse

In Tabelle 7.2 ist aufgetragen zu wie viel Prozent die Stimmungsanalyse bei den im ersten Schritt gefundenen Merkmalen richtig liegt. Dies ist wie bei der Merkmalsextraktion für beide Gruppen, sowie im Durchschnitt dargestellt. Zusätzlich enthält die Tabelle die tatsächliche Anzahl der betrachteten Merkmale je Gruppe. Der Unterschied dieser Werte zwischen den Gruppen ist damit zu erklären, dass es zwischen den beiden Gruppen Abweichungen gab, welche der durch die Merkmalsextraktion gefundenen Merkmale tatsächlich valide Merkmale darstellen. Nur die von der jeweiligen Gruppe

als valide Merkmale erkannten Worte wurde bei der Bewertung der Stimmungsanalyse weiter betrachtet. Die Ergebnisse der Stimmungsanalyse der gefundenen Merkmale sind

Tabelle 7.2: Ergebnisse der Stimmungsanalyse

Kategorie	Gruppe	% richtige Stimmung	Σ bewertete Merkmale
Damenmode	Gruppe 1	85,4%	82
	Gruppe 2	85,9%	79
	Schnitt	85,7%	80,5
Herrenmode	Gruppe 1	93,6%	79
	Gruppe 2	84,6%	78
	Schnitt	88,0%	78,5
Multimedia	Gruppe 1	84,0%	25
	Gruppe 2	86,4%	22
	Schnitt	85,2%	23,5

ähnlich gut wie die der Merkmalsextraktion. Im Schnitt wurde bei allen Kategorien bei über 85% der Merkmale die Stimmung richtig als positiv, negativ oder neutral erkannt. Dies ist sehr konsistent über alle Gruppen und Kategorien hinweg, wobei es einen Ausschlag nach oben bei Gruppe 1 in der Kategorie Herrenmode gibt. Dies bewegt sich jedoch, bezogen auf die Menge der bewerteten Merkmale, im normalen Rahmen der zu erwartender Schwankungen.

Kapitel 8

Fazit und Ausblick

In den folgenden Abschnitten wird das Fazit aus den Ergebnissen der gesamten Arbeit gezogen. Nachfolgend wird ein Ausblick auf mögliche weiterführende Arbeiten gegeben, welche sich aus den Erkenntnissen dieser Arbeit ergeben.

8.1 Fazit

Ziel dieser Arbeit war es zu bewerten, ob es möglich ist automatisiert Merkmale aus nutzergenerierten Onlinebewertungen zu extrahieren und die Stimmung zu diesen Merkmalen zu analysieren.
Dabei wurden zwei Methoden zur Merkmalsextraktion betrachtet und verglichen. Es stellte sich heraus, dass die Methode der Merkmalsextraktion durch häufigste Nomen nur begrenzt verwertbare Ergebnisse liefert. Dies ist bedingt durch die Qualität der Ausgangstexte und somit der hohen Fehlerrate beim Part-of-Speech Tagging und der Rechtschreibkorrektur.
Im Vergleich dazu wurden bei der Merkmalsextraktion mittels Word2Vec deutlich bessere Ergebnisse erzielt. Durch die Erkennung ähnlicher Worte mit Hilfe des Word2Vec Modells wurde die Rechtschreibkorrektur fast vollständig obsolet. Lediglich zur Filterung wurde sie benötigt.
Diese Methode erfordert auch kein Part-of-Speech Tagging, welches sich durch die Qualität der Texte nicht nur als äußerst fehleranfällig, sondern vor allem auch als sehr zeitintensives Verfahren herausgestellt hat.
Um mit den vorliegenden Texten gute Ergebnisse erzielen zu können, werden zur Vorverarbeitung für die Merkmalsextraktion mittels Word2Vec lediglich die Schritte Satzsegmentierung und Tokenisierung in Worte benötigt.
Durch die Reduktion der Vorverarbeitungsschritte werden Fehler vermieden und verhindert, dass Fehler in einem der ersten Schritte in den darauf folgenden Schritten zu weiteren Fehlern führen und sich die initiale Fehlerrate somit je Verarbeitungsschritt potenziert.
Anhand der Ergebnisse der Bewertung der Merkmalsextraktion wird festgestellt, dass in der vorliegenden Domäne *Mode* gute bis sehr gute Resultate erzielt werden konnten und explizite Merkmale, welche nur aus einem Wort bestehen, zuverlässig erkannt werden. Die Ergebnisse mit der Fremddomäne *Multimedia* legen nahe, dass diese Methode jedoch nur innerhalb der Domäne, aus der die Ausgangsliste der manuell annotierten Merkmale stammt, erfolgreich anwendbar ist. Auch falsch geschriebene Merkmale konn-

ten mit dieser Methode aus den Bewertungstexten extrahiert werden.
Die Stimmungsanalyse der gefundenen Merkmale mit Hilfe von Wortlisten mit positiver und negativer Konnotation liefert solide Ergebnisse. Sowohl die Merkmale aus der Domäne Mode, als auch die gefundenen Merkmale aus der Fremddomäne Multimedia konnten zuverlässig in positive, negative und neutrale Merkmale eingeteilt werden.
Auch bei neu hinzugefügten Bewertungen werden durch die Merkmalsextraktion mittels Word2Vec die Merkmale gefunden. Kommen jedoch bei neuen Bewertungen auch neue Merkmale hinzu, welche Worte umfassen, die nicht im Wörterbuch des verwendeten Word2Vec Modells vorkommen, muss das verwendete Word2Vec Modell neu berechnet werden.
Abschließend ist festzustellen, dass sowohl eine gute Merkmalsextraktion als auch eine solide Stimmungserkennung mit vergleichsweise einfachen Methoden realisiert werden konnten. Es ist weder rechenintensives Part-of-Speech Tagging notwendig noch mussten für die Erkennung der Stimmungen komplexe Muster aus Part-of-Speech Tags erstellt werden. Auch ist die Merkmalsextraktion mittels Word2Vec nicht auf tiefe neuronale Netze mit hohem Rechenaufwand gestützt, sondern kommt mit einem flachen neuronalen Netz ohne verdeckte Schichten zu sehr guten Resultaten.
Ein manueller Eingriff ist nur an zwei Stellen notwendig. Zum einen bei der Auswahl eines geeigneten Word2Vec Modells und eventueller zusätzlicher Trainingsdaten. Zum anderen bei der manuellen Annotation von Bewertungen zur Erstellung der Ausgangsliste der Merkmale für die Erweiterung durch Word2Vec. Bei der Stimmungsanalyse können die Listen der Worte mit positiver beziehungsweise negativer Konnotation durch weitere Worte aus der manuellen Annotation erweitert werden. Dies ist aber nicht zwingend notwendig. Werden die Listen durch manuell annotierte Worte erweitert, so ist die Stimmungsanalyse gegenüber falsch geschriebenen Worten robuster.

8.2 Weiterführende Arbeiten

In dieser Arbeit wurde ein zuverlässiger Ansatz gezeigt wie innerhalb einer Domäne Merkmale, welche aus einem Wort bestehen gefunden werden und in positive, negative oder neutrale Merkmale eingeteilt werden.
Mehrwort Merkmale können potentiell auch mit Word2Vec Modellen abgebildet werden. Dabei stellt sich die Frage, ab welcher Datenmenge zuverlässig valide n-Gramme aus Worten errechnet werden können und ob diese zur Erweiterung einer manuell annotierten Merkmalsliste verwendet werden können.
Es wurden in dieser Arbeit auch keine impliziten Merkmale betrachtet. In einer weiterführenden Arbeit kann der Ansatz aus [PE07] genutzt werden, der besagt, dass explizite Merkmale mit den gleichen Meinungsworten bewertet werden wie implizite.
Um die Notwendigkeit eines manuellen Eingriffs bei der Merkmalsextraktion noch weiter zu reduzieren wäre es interessant zu betrachten, wie sich die Menge der manuell annotierten Merkmale in der Ausgangsliste auf die Qualität der Merkmalsextraktion auswirkt.
Anknüpfend an die vorliegende Arbeit ist ein weiterer interessanter Ansatz auch die Gruppierung der Merkmale, sowie die Zusammenfassung der Meinungen zu den Merkmalen je Produkt. Um die Merkmale zu gruppieren ist es denkbar mit Hilfe des Word2Vec Modells Cluster zu berechnen und die Merkmale darüber zu Gruppen zusammenzufassen und zu jeder Merkmalsgruppe den Mittelwert der Stimmungen zu errechnen.

Quellenverzeichnis

Literatur

[BF10] Vertica Bhardwaj und Ann Fairhurst. „Fast fashion: response to changes in the fashion industry“. In: *The International Review of Retail, Distribution and Consumer Research* 20.1 (2010), S. 165–173 (siehe S. 2).

[Bit12] *Trends im E-Commerce, Konsumverhalten beim Online-Shopping*. BITKOM – Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. 2012. URL: http://www.bitkom.org/files/documents/BITKOM_E-Commerce_Studienbericht.pdf (siehe S. 1, 2).

[BK06] Marco Baroni und Adam Kilgarriff. „Large linguistically-processed web corpora for multiple languages“. In: *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*. Association for Computational Linguistics. 2006, S. 87–90 (siehe S. 17).

[BKL09] Steven Bird, Ewan Klein und Edward Loper. *Natural Language Processing with Python*. 1st. O'Reilly Media, Inc., 2009 (siehe S. 12, 15).

[Bla+08] Sasha Blair-Goldensohn u. a. „Building a sentiment summarizer for local service reviews“. In: *WWW Workshop on NLP in the Information Explosion Era*. 2008, S. 14 (siehe S. 7).

[Bra00] Thorsten Brants. „TnT: a statistical part-of-speech tagger“. In: *Proceedings of the sixth conference on Applied natural language processing*. Association for Computational Linguistics. 2000, S. 224–231 (siehe S. 18).

[Car+09] Kai-Uwe Carstensen u. a. *Computerlinguistik und Sprachtechnologie: Eine Einführung*. Hrsg. von Kai-Uwe Carstensen u. a. ISBN 3827420237. Springer, 2009 (siehe S. 14, 15, 17).

[DLY08] Xiaowen Ding, Bing Liu und Philip S Yu. „A holistic lexicon-based approach to opinion mining“. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM. 2008, S. 231–240 (siehe S. 6, 9).

[Ert09] Wolfgang Ertel. *Grundkurs Künstliche Intelligenz: eine praxisorientierte Einführung*. Hrsg. von Florian Mast. Second. ISBN 978-3-8348-0783-0. Wiesbaden: Vieweg + Teubner, 2009, S. 342. URL: http://d-nb.info/994758561 (siehe S. 33).

[FPS96] Usama Fayyad, Gregory Piatetsky-Shapiro und Padhraic Smyth. „The KDD Process for Extracting Useful Knowledge from Volumes of Data“. In: *Commun. ACM* 39.11 (Nov. 1996), S. 27–34. URL: http://doi.acm.org/10.1145/240455.240464 (siehe S. 3).

[GE09] Eugenie Giesbrecht und Stefan Evert. „Is part-of-speech tagging a solved task? an evaluation of pos taggers for the German Web as Corpus“. In: *Proceedings of the Fifth Web as Corpus Workshop*. 2009, S. 27–35 (siehe S. 17, 18).

[HL04] Minqing Hu und Bing Liu. „Mining opinion features in customer reviews“. In: *AAAI*. Bd. 4. 4. 2004, S. 755–760 (siehe S. 7, 20).

[Huf52] David A Huffman. „A method for the construction of minimum redundancy codes“. In: *proc. IRE* 40.9 (1952), S. 1098–1101 (siehe S. 24).

[Jor10] Felipe Jordão Almeida Prado Mattosinho. „Mining Product Opinions and Reviews on the Web“. Masterarbeit. Technische Universität Dresden, Juli 2010. URL: http://www.rn.inf.tu-dresden.de/uploads/studentische_arbeiten/masterarbeit_mattosinho_felipe.pdf (siehe S. 6).

[Kim10] Ina Kimmling. „Opinion Mining - Entwicklung eines Vorgehensmodells“. Masterarbeit. Universität Koblenz-Landau, 2010. URL: http://kola.opus.hbz-nrw.de/volltexte/2010/496/pdf/Masterarbeit_Kimmling.pdf (siehe S. 6).

[Kru+11] Geert-Jan M Kruijff u. a. *NEGRA Homepage*. abgerufen am 23.10.2014. 2011. URL: http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html (siehe S. 17).

[KV12] Walter Kasper und Mihaela Vela. „Monitoring and Summarization of Hotel Reviews“. In: *Information and Communication Technologies in Tourism 2012*. 2012, S. 471–482 (siehe S. 7).

[Mik+13a] Tomas Mikolov u. a. „Distributed representations of words and phrases and their compositionality“. In: *Advances in Neural Information Processing Systems*. 2013, S. 3111–3119 (siehe S. 23, 30).

[Mik+13b] Tomas Mikolov u. a. „Efficient estimation of word representations in vector space“. In: *arXiv preprint arXiv:1301.3781* (2013) (siehe S. 21–25, 28, 30, 32).

[PE07] Ana-Maria Popescu und Orena Etzioni. „Extracting product features and opinions from reviews“. In: *Natural language processing and text mining*. Springer, 2007, S. 9–28 (siehe S. 7, 49).

[PO08] Viktor Pekar und Shiyan Ou. „Discovery of subjective evaluations of product features in hotel reviews“. In: *Journal of Vacation Marketing* 14.2 (2008), S. 145–155.

[Pon+14] Maria Pontiki u. a. „SemEval-2014 Task 4: Aspect Based Sentiment Analysis“. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics und Dublin City University, Aug. 2014, S. 27–35. URL: http://www.aclweb.org/anthology/S14-2004 (siehe S. 7, 36).

[Pon12] Beltrán Borja Fiz Pontiveros. „Opinion Mining from a Large Corpora of Natural Language Reviews“. Masterarbeit. Universitat Politècnica de Catalunya, 2012 (siehe S. 6).

[PTL93] Fernando Pereira, Naftali Tishby und Lillian Lee. „Distributional clustering of English words“. In: *Proceedings of the 31st annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 1993, S. 183–190 (siehe S. 17).

[RQH10] R. Remus, U. Quasthoff und G. Heyer. „SentiWS – a Publicly Available German-language Resource for Sentiment Analysis“. In: *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*. 2010, S. 1168–1171 (siehe S. 11, 43).

[ŘS10] Radim Řehůřek und Petr Sojka. „Software Framework for Topic Modelling with Large Corpora“. English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, 2010, S. 45–50 (siehe S. 12).

[Ton11] Daniel Tondera. „Merkmalsbasierte Stimmungsanalyse aus nutzergenerierten Webinhalten“. Bachelorarbeit. Hochschule der Medien, Stuttgart, 2011 (siehe S. 6, 17, 18).

[TW14] Zhiqiang Toh und Wenting Wang. „DLIREC: Aspect Term Extraction and Term Polarity Classification System“. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics und Dublin City University, Aug. 2014, S. 235–240. URL: http://www.aclweb.org/anthology/S14-2038 (siehe S. 7, 8, 33).

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang A

Anhang zur Vorverarbeitung

A.1 Stuttgart-Tübingen Tagset

Tabelle A.1: Überblick über das Stuttgart-Tübingen Tagset[1]

POS	DESCRIPTION	EXAMPLES
ADJA	attributives Adjektiv	[das] große [Haus]
ADJD	adverbiales oder prädikatives Adjektiv	[er fährt] schnell, [er ist] schnell
ADV	Adverb	schon, bald, doch
APPR	Präposition; Zirkumposition links	in [der Stadt], ohne [mich]
APPRART	Präposition mit Artikel	im [Haus], zur [Sache]
APPO	Postposition	[ihm] zufolge, [der Sache] wegen
APZR	Zirkumposition rechts	[von jetzt] an
ART	bestimmter oder unbestimmter Artikel	der, die, das, ein, eine
CARD	Kardinalzahl	zwei [Männer], [im Jahre] 1994
FM	Fremdsprachliches Material	[Er hat das mit “] A big fish [” übersetzt]
ITJ	Interjektion	mhm, ach, tja
KOUI	unterordnende Konjunktion mit “zu” und Infinitiv	um [zu leben], anstatt [zu fragen]
KOUS	unterordnende Konjunktion mit Satz	weil, dass, damit, wenn, ob
KON	nebenordnende Konjunktion	und, oder, aber
KOKOM	Vergleichskonjunktion	als, wie
NN	normales Nomen	Tisch, Herr, [das] Reisen
NE	Eigennamen	Hans, Hamburg, HSV
PDS	substituierendes Demonstrativpronomen	dieser, jener

PDAT	attribuierendes Demonstrativpronomen	jener [Mensch]
PIS	substituierendes Indefinitpronomen	keiner, viele, man, niemand
PIAT	attribuierendes Indefinitpronomen ohne Determiner	kein [Mensch], irgendein [Glas]
PIDAT	attribuierendes Indefinitpronomen mit Determiner	[ein] wenig [Wasser], [die] beiden [Brüder]
PPER	irreflexives Personalpronomen	ich, er, ihm, mich, dir
PPOSS	substituierendes Possessivpronomen	meins, deiner
PPOSAT	attribuierendes Possessivpronomen	mein [Buch], deine [Mutter]
PRELS	substituierendes Relativpronomen	[der Hund ,] der
PRELAT	attribuierendes Relativpronomen	[der Mann ,] dessen [Hund]
PRF	reflexives Personalpronomen	sich, einander, dich, mir
PWS	substituierendes Interrogativpronomen	wer, was
PWAT	attribuierendes Interrogativpronomen	welche[Farbe], wessen [Hut]
PWAV	adverbiales Interrogativ- oder Relativpronomen	warum, wo, wann, worüber, wobei
PAV	Pronominaladverb	dafür, dabei, deswegen, trotzdem
PTKZU	"zu" vor Infinitiv	zu [gehen]
PTKNEG	Negationspartikel	nicht
PTKVZ	abgetrennter Verbzusatz	[er kommt] an, [er fährt] rad
PTKANT	Antwortpartikel	ja, nein, danke, bitte
PTKA	Partikel bei Adjektiv oder Adverb	am [schönsten], zu [schnell]
TRUNC	Kompositions-Erstglied	An- [und Abreise]
VVFIN	finites Verb, voll	[du] gehst, [wir] kommen [an]
VVIMP	Imperativ, voll	komm [!]
VVINF	Infinitiv, voll	gehen, ankommen
VVIZU	Infinitiv mit "zu", voll	anzukommen, loszulassen
VVPP	Partizip Perfekt, voll	gegangen, angekommen
VAFIN	finites Verb, aux	[du] bist, [wir] werden
VAIMP	Imperativ, aux	sei [ruhig !]
VAINF	Infinitiv, aux	werden, sein
VAPP	Partizip Perfekt, aux	gewesen
VMFIN	finites Verb, modal	dürfen
VMINF	Infinitiv, modal	wollen

VMPP	Partizip Perfekt, modal	gekonnt, [er hat gehen] können
XY	Nichtwort, Sonderzeichen enthaltend	3:7, H2O, D2XW3
$,	Komma	,
$.	Satzbeendende Interpunktion	. ? ! ; :
$(	sonstige Satzzeichen; satzintern	- [,]()

A.2 Perl Script zur Bereinigung der Wikipediadaten

Im folgenden Listing A.1 ist der Quellcode des Scriptes, welches zur Bereinigung der Wikipediadaten verwendet wurde, aufgeführt. Die Änderungen für die Anpassung des Scripts an die deutsche Sprache (ab Zeile 40) sind im darauf folgenden Listing A.2 aufgeführt.

Listing A.1: Perl Script für die Aufbereitung von Wikipediadaten

```
1 #!/usr/bin/perl
2
3 # Program to filter Wikipedia XML dumps to "clean" text consisting only of
      lowercase
4 # letters (a-z, converted from A-Z), and spaces (never consecutive).
5 # All other characters are converted to spaces.  Only text which normally appears
6 # in the web browser is displayed.  Tables are removed.  Image captions are
7 # preserved.  Links are converted to normal text.  Digits are spelled out.
8
9 # Written by Matt Mahoney, June 10, 2006.  This program is released to the public
      domain.
10
11 $/=">";                     # input record separator
12 while (<>) {
13   if (/<text /) {$text=1;}  # remove all but between <text> ... </text>
14   if (/#redirect/i) {$text=0;}  # remove #REDIRECT
15   if ($text) {
16
17     # Remove any text not normally visible
18     if (/<\/text>/) {$text=0;}
19     s/<.*>//;               # remove xml tags
20     s/&amp;/&/g;            # decode URL encoded chars
21     s/&lt;/</g;
22     s/&gt;/>/g;
23     s/<ref[^<]*<\/ref>//g;  # remove references <ref...> ... </ref>
24     s/<[^>]*>//g;           # remove xhtml tags
25     s/\[http:[^] ]*/[/g;    # remove normal url, preserve visible text
26     s/\|thumb//ig;          # remove images links, preserve caption
27     s/\|left//ig;
28     s/\|right//ig;
29     s/\|\d+px//ig;
30     s/\[\[image:[^\[\]]*\|//ig;
31     s/\[\[category:([^|\]]*)[^]]*\]\]/[[$1]]/ig;  # show categories without markup
```

[1] http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html, abgerufen am 19.09.2014

```
32     s/\[\[[a-z\-]*:[^\]]*\]\]//g;  # remove links to other languages
33     s/\[\[[^\|\]]*\|/[[/g;  # remove wiki url, preserve visible text
34     s/{{[^}]*}}//g;         # remove {{icons}} and {tables}
35     s/{[^}]*}//g;
36     s/\[//g;                # remove [ and ]
37     s/\]//g;
38     s/&[^;]*;/ /g;          # remove URL encoded chars
39
40     # convert to lowercase letters and spaces, spell digits
41     $_=" $_ ";
42     tr/A-Z/a-z/;
43     s/0/ zero /g;
44     s/1/ one /g;
45     s/2/ two /g;
46     s/3/ three /g;
47     s/4/ four /g;
48     s/5/ five /g;
49     s/6/ six /g;
50     s/7/ seven /g;
51     s/8/ eight /g;
52     s/9/ nine /g;
53     tr/a-z/ /cs;
54     chop;
55     print $_;
56   }
57 }
```

Listing A.2: geänderte Zeilen des Perl Scriptes für die Aufbereitung von Wikipediadaten

```
40 # convert to lowercase letters and spaces
41     $_=" $_ ";
42   ##### added line with german special characters #####
43     tr/A-ZÄÖÜa-zäöüß.,;!/ /cs;
44   ##### omitted part for use with german wikipedia #####
45   # tr/A-Z/a-z/;
46    # s/0/ zero /g;
47    # s/1/ one /g;
48    # s/2/ two /g;
49    # s/3/ three /g;
50    # s/4/ four /g;
51    # s/5/ five /g;
52    # s/6/ six /g;
53    # s/7/ seven /g;
54    # s/8/ eight /g;
55    # s/9/ nine /g;
56    # tr/a-z/ /cs;
57   ###### end of omitted part
58     chop;
59     print $_;
60   }
61 }
```

Anhang B

Anhang zur Merkmalsextraktion

B.1 Merkmalsextraktion durch häufigste Nomen

Tabelle B.1: 0.5% häufigste Nomen und Eigennamen

Wort	#
farbe	60801
schön	50231
qualität	49183
material	45939
passform	44321
stoff	43723
größe	37461
farben	34294
artikel	29896
preis	24852
schnitt	24031
jacke	21972
groß	16319
schöne	16218
größer	14440
fällt	13337
oberteil	13025
schuh	12572
länge	12415
haut	12318
verarbeitung	12196
sitz	11691
form	11686
sommer	11399

Fortsetzung auf der nächsten Seite

Wort	#
figur	10447
nummer	10431
bild	10109
schönes	9872
frauen	7923
tragekomfort	7303
waschen	7272
pulli	7053
schöner	6832
muster	6069
oberweite	5689
ärmel	5352
größen	5295
fall	5182
abbildung	5056
foto	4699
dünn	4615
tragen	4426
ausschnitt	4337
hingucker	3955
tops	3922
ordnung	3871
zurückgeschickt	3720
sport	3703
winter	3664
katalog	3645
bauch	3635

Fortsetzung auf der nächsten Seite

Wort	#
weiß	3582
hält	3523
spitze	3495
bund	3485
träger	3371
länger	3244
tochter	3220
hosen	3208
große	3205
körper	3192
busen	3129
brust	3122
strand	3092
lässt	3041
tragegefühl	2971
frau	2870
wäre	2818
preis-leistungsverhältnis	2790
klasse	2764
po	2549
geld	2463
hübsch	2394
überbetont	2376
saß	2344
beine	2306
design	2295
wäsche	2287
taille	2257
optik	2249
geschmack	2234
freizeit	2225
fuß	2217
dafür	2156
legierens	2145
schönen	2136
schlafanzug	2117
ok	2103
hüfte	2048
meinung	2040
mann	1985
lieferung	1979
cm	1962

Fortsetzung auf der nächsten Seite

Wort	#
dekolletee	1948
gefällt	1927
wunderschön	1904
muss	1871
kragen	1855
super	1854
großen	1834
sack	1820
hätte	1816
leistung	1814
urlaub	1801
körbchen	1778
beinen	1776
leistungsverhältnis	1762
ware	1753
büro	1740
enttäuscht	1725
höschen	1706
gutes	1698
taschen	1678
trägt	1655
halt	1653
nr..	1630
alltag	1600
süß	1515
beschreibung	1501
anlass	1495
hüften	1474
basic	1472
rücken	1472
füße	1457
kapuze	1439
nähte	1429
spaß	1425
shirley	1382
lässig	1377
slips	1374
bikinihose	1373
coup	1371
dekolleté	1369
auswahl	1348
problem	1336

Fortsetzung auf der nächsten Seite

Wort	#
preis-leistung-verhältnis	1331
hündchen	1301
produkt	1285
nachteil	1284
bein	1282
ausfällt	1266
sterne	1247
farbkombination	1232
stoffqualität	1208
körpergröße	1189
set	1140
sommerhose	1138
hammer	1136
reißverschluss	1129
freundin	1119
bildern	1118
coups	1118
gefühl	1110
fürs	1109
grün	1108
preis-leistung	1101
kombination	1100
kauf	1084
knöpfe	1078
ärmeln	1070
preisleistungsverhältnis	1069
leute	1065
frühling	1065
jahreszeit	1064
türkis	1060
stück	1060
herbst	1057
sachen	1050
geschmackssache	1044
trage	1042
retoure	1042
wäschen	1034
falten	1032
zurückschicken	1031
baumwolle	1029
seite	1023
kürzer	1008

Fortsetzung auf der nächsten Seite

Wort	#
schultern	1008
oberschenkel	1004
erwartungen	1002
passt	993
marke	990
knie	989
möchte	984
hause	970
arbeit	967
(	965
drückt	956
bewertungen	954
armen	945
%	943
aussehen	939
kurzgröße	933
probleme	927
problemzonen	924
wohlfühlen	923
müssen	918
original	918
schwarz	913
sitzen	902
wirklichkeit	882
naht	882
streifen	881
drüber	872
sohle	872
brüste	871
schicke	869
pölsterchen	858
unterwäsche	843
bikinis	840
waden	837
nachteile	836
schuhen	835
sommertage	833
eindruck	828
arme	823
modell	822
bügel	821
damen	820

Fortsetzung auf der nächsten Seite

Wort	#
lang	819
blickfang	804
link	802
anlässe	793
völlig	791
gummizug	791
überein	790
vorstellungen	774
rösser	769
outfit	768
bikinioberteil	749
seiten	747
kleiderschrank	731
überzeugt	728
bestellung	724
verhältnis	724
bäuchlein	719
stiefeln	716
grund	716
eröße	715
leder	714
blusen	708
ruffang	705
anlässen	703
oberteile	694
hals	689
nummern	689
abzug	683
oberschenkeln	682
effekt	680
details	679
größenangabe	672
blau	671
traum	668
gummi	668
okay	664
oberteilen	662
tasche	658
übergang	651
geltung	644
kombi	641
übergangszeit	639

Fortsetzung auf der nächsten Seite

Wort	#
wetter	635
preis/leistung	634
gelegenheiten	623
internet	618
preis/leistungsverhältnis	615
personen	612
partys	609
stern	605
kleidung	605
bush	605
kaufempfehlung	602
petroleum	602
rot	596
überrascht	595
überall	595
manko	591
größere	581
gelegenheit	580
kräftig	579
empfehlen	578
jeansträger	577
preis-/leistungsverhältnis	577

Fortsetzung auf der nächsten Seite

B.2 Ergebnisse der Merkmalsannotation mit *brat*

Tabelle B.2: Liste der manuell annotierten Merkmale

Wort
qualtität
scnitt
materila
ausschnitt
farbkombinationen
farbkombi
preis-, leistungsverhältnis
gewicht
abnäher/nähte
trageigenschaft
rücken
fleece-stoff
knallfarbe
strickstoff
optik
farbton
tragegefühl
ärmel
farbverlauf
preis/ leistungsverhältnis
kurz-größe
beinausschnitt
preil-leistungs-verhältnis
design
tragekomfort
preis - leistung
ausführung
farbe
stoffmaterial
shnitt
sitz
herbstfarbe
preis- leistung
oberteil
preislleistungsverhältnis
Fortsetzung auf der nächsten Seite

Wort
farbgestaltung
preis-leistungsverhälnis
nadelstreifen
verschluss
qualitätseindruck
weste
größe
look
formgebung
komfort
pastellfarben
qualitat
spitze
gürtellänge
reißverschluß
preis- leistungsverhältnis
körperlänge
dekolltè
raffungen
hosenstoff
aufmachung
armausschnitt
trageeigenschaft
stoffqualität
sommerfarben
waschung
rippmaterial
beinlänge
verarbeitung
preis-leistung-verhältnis
stoff/
qualität
schriftzug
bündchen
tragbarkeit
trageigenschaften
druck
Fortsetzung auf der nächsten Seite

Wort
quatlität
höhe
preis-leistungsverhältnis
passform
nähte
stooffqualität
trage-/körpergefühl
matrial
preis leistung
schnittführung
quallität
musterprint
laenge
gesäßtaschen
farbqualität
geruch
dekollete
größenangaben
quali
rückenteil
pasform
träger
preis-leistungsverhältniss
grösse
absatz
passvorm
futter
preis/leistungsverhätniss
farbkombination
normalpreis
farben
rüschen
farbauswahl
passforn
brustbereich
baumwolle
verarbeiterung
sitzform
halt
braunton
materia
rückenprint
Fortsetzung auf der nächsten Seite

Wort
preisleistungsverhältniss
cups
schnitt
preis leistung verhältnis
jeansstoff
schnittes
formkraft
passfrom
verarbeitungsqualität
abnäher
spitzenstickerei
laufkomfor
farbvariante
blazer
bund
stretch-material
preiß
ärmeln
farbgebeung
obermaterial
preis leistungsverhältnis
schnittform
material
frabe
oaulität
materal
silhouette
ärmellänge
leistung
farbenkombination
materialqualität
farbvarianten
qualitaet
trägerkombinationen
qaulität
sohle
passkomfort
schnit
geld
spitzenrand
preisleistungsverhältnis
farbkontrast
Fortsetzung auf der nächsten Seite

Wort
sommer-stoff
qulität
brusteinsätze
länge
stege
details
stoffart
print
größenabgabe
seitenstreifen
stof
stoffes
tragqualität
qualiltät
aufdruck
modell
fleecematerial
verabeitung
gewebe
größenangabe
baumwollstoff
dekoltee
tragekonfort
naht
preisleistungverhältnis
schnittführun
preis-/leistungsverhältnis
preis/leistungverhältniss
preis/leistungsverhältnis
hosenlänge
rückenausschnitt
form
maschenbild
absatzhöhe
färbe
bügel
büstier
preis-leistung
preis -leistung
preis/leistung
preisleistung
muster

Fortsetzung auf der nächsten Seite

Wort
trageekomfort
paßform
schleife
trageeigenschaften
reisverschluss
größen
preis–leistungsverhältnis
aussehen
musterung
farbgebung
leibhöhe
preis-leistungs-verhältnis
preis - leistungsverhältnis
leder
tragekompfort
blumenmuster
kragen
breite
preis
stoff
preis-leistungs-verhältnisse
dekolltee

Fortsetzung auf der nächsten Seite

B.3 Wortversionen

Tabelle B.3: Versionen des Wortes *Preisleistungsverhältnis*

Wort	lässt sich korrigieren
preil-leistungs-verhältnis	Nein
preislleistungsverhältnis	Ja
preis-leistungsverhälnis	Ja
preis-leistung-verhältnis	Ja
preis-leistungsverhältnis	Ja
preis-leistungsverhältniss	Ja
preis/leistungsverhätniss	Nein
preisleistungsverhältniss	Ja
preisleistungverhältnis	Ja
preis-/leistungsverhältnis	Ja
preis/leistungverhältniss	Nein
preis/leistungsverhältnis	Ja
preis–leistungsverhältnis	Ja
preis-leistungs-verhältnis	Ja
preis-leistungs-verhältnisse	Nein
preis-leitungs-verhältnis	Nein
preis-/leistungsverhältniss	Nein
preis-leistungs-verhältniss	Nein
preis-/leistungverhältnis	Nein
preis/leistungs-verhältnis	Ja

Tabelle B.4: Versionen des Wortes *Qualität*

Wort	lässt sich korrigieren
qualtität	Ja
qualitat	Nein
qualitaet	Nein
quatlität	Ja
quali	Nein
qaulität	Ja
qualiltät	Ja
kwalitet	Nein
quaität	Ja
qulität	Ja
qualiät	Ja
quallität	Ja
qalitet	Nein
qualtität	Ja
qalität	Ja
qullität	Ja
quallitat	Nein
qaulität	Ja
quälität	Ja
qualitöt	Ja

B.4 Semantische und syntaktische Fragen

Im Folgenden findet sich eine Liste der Wortpaare aus welchen die semantischen und syntaktischen Fragen zusammengesetzt sind. Für jeden Bereich werden die Wortpaare zu Fragen an das Word2Vec Modell kombiniert, so dass für jeden Bereich 420 Fragen enstehen. Jeder Bereich enthält 21 Wortpaare.

: Gegenteil
gut schlecht
groß klein
hell dunkel
stark schwach
lang kurz
viel wenig
hart weich
gemütlich ungemütlich
dick dünn
leicht schwer
sicher unsicher
angenehm unangenehm
angemessen unangemessen
gerechtfertigt ungerechtfertigt
erwartet unerwartet
glücklich unglücklich
bekannt unbekannt
wahrscheinlich unwahrscheinlich
logisch unlogisch
verständlich unverständlich
bewusst unbewusst
: Perfekt erste Person Singular
bestellt bestelle
gesagt sage
gesucht suche
gedacht denke
gemeint meine
gekauft kaufe
probiert probiere
gezahlt zahle
gegangen gehe
gelaufen laufe
gefühlt fühle
gewusst weiß
gewollt will
getragen trage
gefunden finde
gemocht mag
gekonnt kann
getestet teste
gewaschen wasche
gehört höre
gestanden stehe
: Präteritum dritte Person Singular
saß sitzt
wirkte wirkt
passte passt
stand steht
ging geht
hatte hat
probierte probiert
kratzte kratzt
juckte juckt
lief läuft
bekam bekommt
fiel fällt
gefiel gefällt
mochte mag
dachte denkt
wärmte wärmt
sah sieht
fühlte fühlt
hörte hört
wollte will
glänzte glänzt
: Steigerung
hell heller
dunkel dunkler
gut besser
schlecht schlechter
dick dicker
dünn dünner
günstig günstiger
teuer teurer
groß größer

klein kleiner
kalt kälter
schlank schlanker
eng enger
einfach einfacher
kurz kürzer
jung jünger
laut lauter
lang länger
schwer schwerer
schnell schneller
leicht leichter
: Plural Nomen
hemd hemden
shirt shirts
hose hosen
schuh schuhe
rock röcke
absatz absätze
oberteil oberteile
frau frauen
mann männer
jacke jacken
tasche taschen
arm arme
bein beine
paket pakete
bluse blusen
mantel mäntel
farbe farben
größe größen
preis preise
kind kinder
person personen
: Geschlecht
männer frauen
herren damen
ihm ihr
er sie
freund freundin
mann frau
sohn tochter
opa oma
vater mutter
enkel enkelin
bruder schwester
seine ihre
onkel tante
junge mädchen
neffe nichte
großvater großmutter
jungs mädels
opas omas
väter mütter
bräutigam braut
stiefvater stiefmutter
: Superlativ
hellste hell
dunkelste dunkel
schlechteste schlecht
beste gut
günstigste günstig
teuerste teuer
größte groß
kleinste klein
kälteste kalt
schnellste schnell
höchste hoch
längste lang
tiefste tief
glücklichste glücklich
älteste alt
jüngste jung
einfachste einfach
kürzeste kurz
langsamste langsam
seltsamste seltsam
stärkste stark
: Präteritum erste Person Singular
kenne kannte
will wollte
kaufe kaufte
trage trug
bekomme bekam
laufe lief
probiere probierte
habe hatte
gehe ging
stehe stand
passe passte
wirke wirkte
denke dachte
finde fand
mag mochte

bestelle bestellte
kann konnte
träume träumte
wünsche wünschte
hoffe hoffte
glaube glaubte

B.5 Erweiterte Merkmalsliste

Tabelle B.5: Erweiterte Merkmalsliste mit einer minimalen Kosinus Distanz von 0.65

Wort
qualtität
reißverschluss
scnitt
materila
ausschnitt
farbkombinationen
farbkombi
preis-, leistungsverhältnis
gewicht
abnäher/nähte
trageigenschaft
rücken
fleece-stoff
knallfarbe
strickstoff
optik
farbton
tragegefühl
ärmel
farbverlauf
preis/ leistungsverhältnis
kurz-größe
beinausschnitt
preil-leistungs-verhältnis
design
tragekomfort
preis - leistung
ausführung
farbe
stoffmaterial
shnitt
sitz
herbstfarbe
preis- leistung
Fortsetzung auf der nächsten Seite

Wort
oberteil
preislleistungsverhältnis
farbgestaltung
preis-leistungsverhälnis
nadelstreifen
verschluss
qualitätseindruck
weste
größe
look
formgebung
komfort
pastellfarben
qualitat
spitze
gürtellänge
reißverschluß
preis- leistungsverhältnis
körperlänge
dekolltè
raffungen
hosenstoff
aufmachung
armausschnitt
trageeigenschaft
stoffqualität
sommerfarben
waschung
rippmaterial
beinlänge
verarbeitung
preis-leistung-verhältnis
stoff/
qualität
schriftzug
bündchen
tragbarkeit
Fortsetzung auf der nächsten Seite

Wort
trageigenschaften
druck
quatlität
höhe
preis-leistungsverhältnis
passform
nähte
stooffqualität
trage-/körpergefühl
matrial
preis leistung
schnittführung
quallität
musterprint
laenge
gesäßtaschen
farbqualität
geruch
dekollete
größenangaben
quali
rückenteil
pasform
träger
preis-leistungsverhältniss
grösse
absatz
passvorm
futter
preis/leistungsverhätniss
farbkombination
normalpreis
farben
rüschen
farbauswahl
passforn
brustbereich
baumwolle
verarbeiterung
sitzform
halt
braunton
Fortsetzung auf der nächsten Seite

Wort
materia
rückenprint
preisleistungsverhältniss
cups
schnitt
preis leistung verhältnis
jeansstoff
schnittes
formkraft
passfrom
verarbeitungsqualität
abnäher
spitzenstickerei
laufkomfor
farbvariante
blazer
bund
stretch-material
preiß
ärmeln
farbgebeung
obermaterial
preis leistungsverhältnis
schnittform
material
frabe
oaulität
materal
silhouette
ärmellänge
leistung
farbenkombination
materialqualität
farbvarianten
qualitaet
trägerkombinationen
qaulität
sohle
passkomfort
schnit
geld
spitzenrand
Fortsetzung auf der nächsten Seite

Wort
preisleistungsverhältnis
farbkontrast
sommer-stoff
qulität
brusteinsätze
länge
stege
details
stoffart
print
größenabgabe
seitenstreifen
stof
stoffes
tragqualität
qualiltät
aufdruck
modell
fleecematerial
verabeitung
gewebe
größenangabe
baumwollstoff
dekoltee
tragekonfort
naht
preisleistungverhältnis
schnittführun
preis-/leistungsverhältnis
preis/leistungverhältniss
preis/leistungsverhältnis
hosenlänge
rückenausschnitt
form
maschenbild
absatzhöhe
färbe
bügel
büstier
preis-leistung
preis -leistung
preis/leistung
Fortsetzung auf der nächsten Seite

Wort
preisleistung
muster
trageekomfort
paßform
schleife
trageeigenschaften
reisverschluss
größen
preis–leistungsverhältnis
aussehen
musterung
farbgebung
leibhöhe
preis-leistungs-verhältnis
preis - leistungsverhältnis
leder
tragekompfort
blumenmuster
kragen
breite
preis
stoff
preis-leistungs-verhältnisse
dekolltee
chino-schnitt
prei-leistung
preis-leitungs-verhältnis
targekomfort
leistungsverhaeltnis
materieal
tragekomport
qüalität
matrieal
blütenmuster
dekollteé
tragecomfort
spitzen-bh
n-länge
verwaschung
stoffanteil
tragekommfort
pastellton
Fortsetzung auf der nächsten Seite

Wort
innenbein
preis-/
preis-/leistungsverhältniss
preis-/leistung
-preis
farbe-
türkisfarben
qualtät
1a-passform
gewöhnungs
strechmaterial
bh-schalen
holzringe
verleit
superchic
qualitet
beinende
baumwolle-qualität
frühlingsfarbe
trage-eigenschaften
qualietät
wascheigenschaften
trendfarben
rüschchen
quitscht
gummieinsatz
strickmaterial
paaform
used-optik
beinabschluss
pssform
verlängerer
hautgefühl
brustmitte
dekolletée
verwaschenen
trageangenehme
normallänge
bundhöhe
dieträger
preisverhältnis
fliesende
Fortsetzung auf der nächsten Seite

Wort
innenärmel
frühlingsfarben
top-ware
musterkombination
qualtiät
dekoletté
trage-komfort
dunklerem
glanzstreifen
qwalität
schweißfüsse
knitterfreien
häkelspitze
decolltee
kompenierbar
normal-größe
preis-leistungs-verhältniss
strandteil
materialzusammenstellung
beinabschluß
querstreifen
sahnestück
supersüße
ärmelausschnitte
ärmelabschlüssen
verarbeitug
grobmaschiger
glitzeroptik
qalitet
trägerlösung
decolté
flegeleicht
quallitat
decoltee
tragform
schönne
mittelbraun
blütendruck
quatität
decollte
schadstoffgeprüft
kordelzug
Fortsetzung auf der nächsten Seite

Wort
feincord
preisleistungs
topmodische
blickdichtem
gutequalität
taille/hüfte
kühlendes
stoffkombination
grasgrün
dekolleté
dekolletè
kwalität
decollté
abschlussbündchen
jeansoptik
geschenkidee
baumwollqualität
super-gut
business-hose
abgesteppten
paßgenauigkeit
armausschnitte
reissverschlusses
speckpölsterchen
bünchen
zurechtzupfen
n-größen
passfprm
top-passform
leuchtene
k-grösse
rückenverzierung
gerüschte
netzmuster
shirtbluse
spitzenborte
türkisblau
meliert
normalgrösse
preis/qualität
klettverschluß
geschitten
Fortsetzung auf der nächsten Seite

Wort
sommerlaune
spitzenstoff
laufkomfort
leistungsverhältniss
dekolté
dekoltet
tragequalität
qualitätsanmutung
weit/zu
jeansqualität
leistungsverhätnis
darübertragen
reissverschluss
ärmelende
matterial
supersüß
dekolte´
sportlich-
preis-und
ärmelabschlüsse
farbstellung
traumfarbe
jeans-material
jeansmaterial
elastan-anteil
blümchenmuster
preis/
herbstfarben
dekolteé
kvalitet
verareitung
schnürre
angenemes
passgerechte
glanzoptik
stoff-qualität
eindurck
trageeingenschaften
bretthart
vearbeitung
smaragdgrün
lang-größe
Fortsetzung auf der nächsten Seite

Wort
enttäusch
tagekomfort
begueme
stretch-qualität
verarbeitung-
lässigem
dekollete´
vernähung
dekoletee
hellbeige
alltagsshirt
kirschrot
spitzenmaterial
ziertaschen
stretchige
trägerform
armausschnitten
hosenende
fliederfarben
qwalitet
printmuster
spitzeneinsatz
mausgrau
jeans-stoff
dekollete‘
hünsch
dekolletee
dekokte
saumabschluss
wünderschöne
schlankmachende
preisleitung
aufläd
elegannt
dekoltée
sommermode
gefeld
materail
-leistungsverhältnis
asymetrische
grünton
l-grösse
Fortsetzung auf der nächsten Seite

Wort
hautfreundliche
sandfarben
langgrösse
sporttop
anschmiegsame
verziehrung
his-logo
cremeweiß
qalität
qualitiät
rostfarben
kragenbereich
farbbrillianz
zeichent
rippenstruktur
l-größe
ärmelbund
-verhältnis
blumenprint
blitzversand
tragkomfort
passorm
dehnbund
doppellagig
tunikabluse
matrerial
bh-rand
altrosa
gräulicher
ferarbeitung
preis-/leistungverhältnis
frauenfuß
quwalität
material-
fröhlichere
männerunterwäsche
preis/leistungs
trendfarbe
trägerhalter
qualiät
brauntöne
ärmellösung
Fortsetzung auf der nächsten Seite

Wort
sommerfeeling
viskosematerial
blumendruck
schlabert
pobereich
n-größe
einsteigersmartphone
speckpolster
bauch/hüfte
stretchqualität
außen-
grundfarbton
modestück
potaschen
dekollette
kurzgroesse
schlappern
gesmokten
beigeton
tragegfühl
qualitätund
reißverschlusstasche
graublau
dekoleté
quaität
knitterfreie
ärmelausschnitt
qallität
ziergürtel
dekolette
qualitität
angesagtem
preis-qualitätsverhältnis
abreibt
angenähten
sockenbund
dekolletté
perlmuttknöpfe
qualität/
qualität-
stoffzusammensetzung
leistungverhältnis
Fortsetzung auf der nächsten Seite

Wort
mittelnaht
kurzgrösse
stretchbund
oualität
brauntönen
strechigen
materialverarbeitung
tragekonform
decolleté
dekolete
farben-
reisverschluß
tragekomfor
l-länge
brustansatz
hammergeile
alltags-
spitzenbesatz
kwalitet
wäscheset
spitzenkante
glitzerdruck
trageform
sommerhighlight
tragefähigkeit
stoffoberfläche
dekollte
rundhalsausschnitt
hellerer
baumwoll-qualität
t-shirt-stoff
verabreitung
marerial
spitzeneinsätze
denimstoff
frühlingsbluse
dekoltè
angehnehmer
zippelt
normal-länge
lachsfarben
partyabende
Fortsetzung auf der nächsten Seite

Wort
läsig
halsauschnitt
dekolltée
absteht
knuddelt
markenqulität
feinstrick
zopfmuster
dekollté
längsstreifen
k-größe
passfor
quälität
sommerleichte
idealle
preis-leistungs
passfom
größem
qualitöt
sommerfarbe
sommerfrisch
körpergrösse
netzeinsatz
verarbetung
zierknöpfe
qullität
schnürrung

Fortsetzung auf der nächsten Seite